Annals of Epidemiology 24 (2014) 673–680

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Nomogram for sample size calculation on a straightforward basis for the kappa statistic

Hyunsook Hong MS [a,b], Yunhee Choi PhD [a,*], Seokyung Hahn PhD [a,c], Sue Kyung Park PhD [d,e,f], Byung-Joo Park PhD [d]

[a] Division of Medical Statistics, Medical Research Collaborating Center, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Korea
[b] Department of Biostatistics, Seoul National University School of Public Health, Seoul, Korea
[c] Department of Medicine, Seoul National University College of Medicine, Seoul, Korea
[d] Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Korea
[e] Department of Biomedical Science, Seoul National University Graduate School, Seoul, Korea
[f] Cancer Research Institute, Seoul National University College of Medicine, Seoul, Korea

ARTICLE INFO

Article history:
Received 19 February 2014
Accepted 26 June 2014
Available online 3 July 2014

Keywords:
Sample size
Kappa statistic
Proportion of agreement
Nomogram

ABSTRACT

Purpose: Kappa is a widely used measure of agreement. However, it may not be straightforward in some situation such as sample size calculation due to the kappa paradox: high agreement but low kappa. Hence, it seems reasonable in sample size calculation that the level of agreement under a certain marginal prevalence is considered in terms of a simple proportion of agreement rather than a kappa value. Therefore, sample size formulae and nomograms using a simple proportion of agreement rather than a kappa under certain marginal prevalences are proposed.
Methods: A sample size formula was derived using the kappa statistic under the common correlation model and goodness-of-fit statistic. The nomogram for the sample size formula was developed using SAS 9.3.
Results: The sample size formulae using a simple proportion of agreement instead of a kappa statistic and nomograms to eliminate the inconvenience of using a mathematical formula were produced.
Conclusions: A nomogram for sample size calculation with a simple proportion of agreement should be useful in the planning stages when the focus of interest is on testing the hypothesis of interobserver agreement involving two raters and nominal outcome measures.

Introduction

Agreement in diagnostic measurements such as inter- and intraobserver agreement, reproducibility with time, or the accuracies of some diagnostic tests compared with a gold standard involves a variety of issues. The proper statistics should be applied based on the type of diagnosis (continuous or categorical), diagnostic problems (consistency or absolute agreement), or data structure (existence of correlations or not).

The most general statistical measure has been Cohen's kappa for categorical diagnostic measurements [1]. It is well known that the simple proportion of agreement (calculated over all paired measurements) is not a proper measure because it consists of not only the true agreement but also agreement that would be expected purely by chance. The agreement by chance depends on the marginal prevalence of ratings. For example, when the diagnostic measurements are disease positive and disease negative, the marginal prevalence is the proportion of each disease type. As the disease type is more equally distributed, the agreement by chance should increase. The kappa statistic is the measure adjusted to reflect the marginal prevalence, so interpreted as the agreement beyond what is expected by chance.

Although kappa is a widely used measure of agreement, it may not be straightforward in some situation such as sample size calculation [2]. For example, in a given sample size calculation, the expected agreement may be assumed to be $\kappa = 0.61$. In general, although $\kappa = 0.61$ is known to be substantial agreement, there are many possible combinations of proportions of agreement and marginal prevalence that could produce $\kappa = 0.61$: from roughly 80.5% to 99.9% agreement depending on the marginal prevalence. The difference in the proportion of agreement between 80.5% and 99% could produce different interpretations in real applications.

Source of funding: None.
* Corresponding author. Medical Research Collaborating Center, Seoul National University Hospital, Daehak-ro, Jongno-gu, Seoul 110-744, Korea. Tel.: +82 2 740 8940; fax: 82 2 766 8172.
E-mail address: yhchoi@snuh.org (Y. Choi).

1047-2797/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.annepidem.2014.06.097

Also, in an another example [3], a null hypothesis kappa value was set at 0.4, representing poor agreement, requiring an analysis of 405 subjects to detect a statistically significantly higher kappa coefficient value of at least 0.61 (representing substantial agreement) at a power of 80% with a marginal prevalence of positive rating 10%. Under the marginal prevalence, the agreement proportions corresponding to $\kappa = 0.4$ and $\kappa = 0.61$ are 90% and 94%, respectively. Only 4% more agreement produces enhancement from poor to substantial agreement in terms of kappa coefficient. If 4% more agreement can be negligible difference in the diagnosis, then the rationale of the study may need to be reconsidered. This phenomenon is related to the kappa paradox: high agreement but low kappa. Hence, it seems reasonable in sample size calculation that the level of agreement under a certain marginal prevalence is considered in terms of a simple proportion of agreement rather than a kappa value. Fortunately, the relationship between a kappa coefficient and a proportion of agreement under a certain marginal prevalence has already been derived [4]. Therefore, we describe here when a proportion of agreement can be a more straightforward basis for a sample size calculation than a kappa value. We have also provided a sample size formula in terms of a proportion of agreement using an existing sample size formula for kappa.

A few methods of sample size calculation for an interobserver agreement study have been suggested (5, 6, 7, and 8). They usually present both a sample size formula and table with sample size under specific conditions. However, the tables can only be restrictively applied because they present only sample sizes under specific conditions. Hence, we provide nomograms to cover a wider range of conditions for sample size calculation.

In this study, a kappa statistic under the common correlation model [4] is used to relate a kappa with a simple proportion of agreement under a certain marginal prevalence. The relationship is applied to sample size calculation based on the goodness-of-fit test [6]. Finally, sample size formulae and nomograms using a simple proportion of agreement and marginal prevalences are provided.

Methods

The kappa statistic

Two-category nominal data may be arranged in four cells of a 2-by-2 table as shown in Table 1. n_1 or n_4 indicates the number of cases in which both raters made the same diagnosis of positive or negative, respectively. Where two raters disagree, n_2 indicates the number of cases where rater 1 made a negative diagnosis, whereas rater 2 made a positive diagnosis. n_3 indicates the opposite form of disagreement. Overall, the percent agreement is defined as the proportion in which the two raters give the same diagnosis, $(n_1 + n_4)/4$.

As mentioned before, agreement by chance alone is included in this overall percent agreement and it depends on (c_1, c_2, r_1, and r_2). The measure of the chance-corrected agreement, Cohen's kappa is defined as follows [9].

Table 1
Data layout in a study with two raters and binary outcome measures

	Rater 1		Total
	Positive	Negative	
Rater 2			
Positive	n_1	n_2	r_1
Negative	n_3	n_4	r_2
Total	c_1	c_2	n

$$k = \frac{\text{observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}} = \frac{p_0 - p_e}{1 - p_e}$$

where $p_0 = \frac{n_1 + n_4}{n}$, $p_e = \frac{r_1 c_1 + r_2 c_2}{n^2}$

p_0 denotes overall percent agreement. p_e denotes the agreement expected by chance, which means that two raters classifying subjects quite independently will agree by chance on a predictable proportion of cases. Kappa is obtained from the overall percent agreement and expected agreement by chance and is a measure of the agreement beyond that expected by chance. The range of possible value of kappa is from −1 to 1. $\kappa = 1$ represents perfect agreement, indicating that the rates agree in their diagnosis of every subject. $\kappa = 0$ indicates agreement no better than that expected by chance, as if the raters had simply "guessed" every rating. A positive (or negative) kappa would indicate agreement better (or worse) than that expected by chance. Landis and Koch [10] have proposed the following as standards for the strength of agreement for the kappa statistic: $\kappa \le 0$: poor, 0.01–0.2: slight, 0.21–0.40: fair, 0.41–0.60: moderate, 0.61–0.80: substantial, and 0.81–1.0: almost perfect.

From the previously mentioned kappa statistics, under a constant p_0, better agreement is obtained as p_e decreases. As the marginal prevalence of positive rating gets close to 0.5, p_e becomes smaller and produces a larger kappa statistic.

The relationship between a kappa and simple proportion of agreement under a given marginal prevalence

Mak [4] proposed a kappa statistic under the common correlation model assuming that the correlation between two raters is constant for all subjects (or images) where the correlation is the same as a kappa value [4]. Let x_{ij} denotes positive or negative diagnoses for the i-th subject (or image) by the j-th rater, where $i = 1, 2, \ldots, n$ and $j = 1, 2$. Assume that Pr (x_{ij} = positive = 1) = π_1 and, Pr (x_{ij} = negative = 0) = $\pi_2 = (1 - \pi_1)$. If the correlation between any pair x_{i1}, x_{i2} has the same value κ, then we have

$$P_1(\kappa) = \Pr(x_{i1} = x_{i2} = 1) = \pi_1^2 + \kappa\pi_1\pi_2$$
$$P_1(\kappa) = \Pr(x_{i1} = 1 \text{ and } x_{i2} = 0) = \Pr(x_{i1} = 0 \text{ and } x_{i2} = 1) = \pi_1\pi_2(1-\kappa)$$
$$P_1(\kappa) = \Pr(x_{i1} = x_{i2} = 0) = \pi_2^2 + \kappa\pi_1\pi_2$$

which denote that the proportion of agreement $p_0 = P_1(\kappa) + P_3(\kappa) = 1 - 2P_2(\kappa) = 1 - 2\pi_1\pi_2(1-\kappa)$, and the expected proportion of agreement $p_e = 1 - 2\pi_1\pi_2$ because $\kappa = 0$. Estimates of π_1, p_0, and p_e are given by, [11]

$$\widehat{\pi}_1 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij} = \frac{c_1 + r_1}{2n}, \text{ and } \widehat{\pi}_2 = 1 - \widehat{\pi}_1$$
$$\widehat{p}_o = \frac{n_1 + n_4}{n}$$
$$\widehat{p}_e = 1 - 2\widehat{\pi}_1\widehat{\pi}_2$$

where c_1, r_1, n_1, and n_4 are in Table 1

The resulting estimator of κ is given by

$$\widehat{\kappa} = \frac{\widehat{p}_o - \widehat{p}_e}{1 - \widehat{p}_e} = \frac{\widehat{p}_o - (1 - 2\widehat{\pi}_1\widehat{\pi}_2)}{2\widehat{\pi}_1\widehat{\pi}_2} = 1 - \frac{1 - \widehat{p}_o}{2\widehat{\pi}_1\widehat{\pi}_2} \quad (1)$$

$$\Leftrightarrow \quad \widehat{p}_o = 1 - 2\widehat{\pi}_1\widehat{\pi}_2(1 - \widehat{\kappa}) = 1 - \left(1 - \widehat{\pi}_1^2 - \widehat{\pi}_2^2\right)(1 - \widehat{\kappa})$$

Equation 1 shows that the kappa statistic is a function of a marginal prevalence and a proportion of agreement. It also shows