ELSEVIER

Journal of Clinical Epidemiology 65 (2012) 778−784

Journal of Clinical Epidemiology

A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes

Michael A. Rotondi[a,*], Allan Donner[b]

[a]School of Kinesiology and Health Sciences, York University, Room 364, Norman Bethune College, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3
[b]Department of Epidemiology and Biostatistics, University of Western Ontario, Room K201, Kresge Building, London, Ontario, Canada N6A 5C1

Accepted 30 October 2011; Published online 4 May 2012

Abstract

Objective: Studies measuring interobserver agreement (reliability) are common in clinical practice, yet discussion of appropriate sample size estimation techniques is minimal as compared with clinical trials. The authors propose a sample size estimation technique to achieve a prespecified lower and upper limit for a confidence interval for the κ coefficient in studies of interobserver agreement.

Study Design and Setting: The proposed technique can be used to design a study measuring interobserver agreement with any number of outcomes and any number of raters. Potential application areas include: pathology, psychiatry, dentistry, and physical therapy.

Results: This technique is illustrated using two examples. The first considers a pilot study in oral radiology, whose authors studied the reliability of the mandibular cortical index as measured by three dental professionals. The second example examines the level of interobserver agreement among four nurses with respect to five triage levels used in the Canadian Triage and Acuity Scale.

Conclusion: This method should be useful in the planning stages of an interobserver agreement study in which the investigator would like to obtain a prespecified level of precision in the estimation of κ. An R software package (R Foundation for Statistical Computing, Vienna, Austria), kappaSize is also provided that implements this method.

Keywords: Kappa; Agreement; Sample size estimation; Reliability; Experimental design

1. Introduction

Studies of interobserver agreement play an important role in medical research, with examples ranging from the evaluation of digitized cervical images in the diagnosis of human papillomavirus, infection states [1], to the appropriate classification of strokes in the Physicians' Health study [2], as well as numerous other applications in the areas of psychology and pathology. Nonetheless, aside from specific subject matter areas (e.g., physical therapy [3] and nutritional screening [4]), the literature on sample size estimation for studies of interobserver agreement is limited.

Analytical approaches for this problem tend to focus on the kappa (κ) family of statistics, given their common use in practice [5]. We focus our attention here on the intraclass κ coefficient [6]. This version of κ is obtained by calculating the intraclass correlation coefficient resulting from a one-way analysis of variance and applying it to binary data. Under this model, the goodness-of-fit (GOF) statistic [7] may be used for hypothesis testing and confidence interval (CI) construction for κ.

Hypothesis testing and CI construction for the intraclass kappa coefficient are well established in the case of binary and multinomial outcomes [7−9]. However, the corresponding literature on sample size estimation is relatively sparse.

Although most of the literature approaches sample size estimation from a hypothesis testing perspective, there has been increasing attention on the role of CI construction in the early stages of experimental design [10]. In contrast to the hypothesis testing approach, the CI approach allows the investigators to design their study with the purpose of obtaining a prespecified level of precision about the point estimate of κ. For example, given a fixed number of raters n, the required number of subjects, N may be calculated so that the expected lower limit of a 95% confidence limit for κ is no less than a specified threshold value κ_L, while simultaneously specifying the expected upper limit, κ_U.

Our objective here is thus twofold: (1) to generalize the CI approach for sample size estimation to the case of multinomial outcomes and multiple raters and (2) to illustrate the use of this approach in the context of two concrete

* Corresponding author. Tel.: +416-736-2100 ext. 22462; fax: +416-736-5774.
E-mail address: mrotondi@yorku.ca (M.A. Rotondi).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2011.10.019

What is new?

- A statistical method to design an interobserver agreement study with a certain degree of precision in the estimation of κ is presented.
- The method was previously only available for interobserver agreement studies with a binary outcome. The technique can now accommodate any number of raters, as well as multiple outcome categories.
- An R software package, kappaSize is available to implement this sample size estimation technique.
- Studies measuring interobserver agreement should consider the desired level of confidence in the estimation of the kappa statistic in the design phase of a planned study.

examples. The first objective represents a direct extension of recent work [11,12], which presents a CI approach to sample size estimation for κ in the case of binary outcomes and two or more raters respectively. The second objective is discussed in the context of examples in oral radiology and emergency medicine. The authors also provide an R software package [28], kappaSize that implements the described procedures.

2. Methods

To ensure consistency with the literature, we adapt the following notation and derivations [8]. Let n represent the number of raters who will rate a sample of N subjects independently into k mutually exclusive categories. Let X_{ij} denote the number of ratings on subject $i(i=1,\ldots,N)$that fall into category $j(j=1,\ldots,k)$ and denote the probability of a rating falling into each category by $\pi_1,\pi_2,\ldots,\pi_k$, respectively, where $\sum_{j=1}^{k}\pi_j=1$ and $\sum_{j=1}^{k}X_{ij}=n$. Note that we assume marginal homogeneity across all raters, that is, the probability of a rating falling into a particular category is constant for each of the raters. The model underlying this assumption is most appropriate when the main emphasis is directed at the reliability of the measurement process rather than differences among raters [6]. We also note the argument by Zwick [13] that ''if one rejects the assumption of marginal homogeneity, one need go no further,'' as the degree of disagreement between raters may be expressed using their marginal distributions, that is, without the kappa statistic. A formal test of this assumption in the case of two raters can be obtained using McNemar's test [14] in the case of a binary outcome, or the Stuart–Maxwell test [15,16] in the case of three outcome categories. Additional details regarding the assumption of marginal homogeneity can be found in Agresti [17] (Chapter 10).

In the case of the GOF procedure, all disagreements are treated equally, that is, no partial credit is given for classifying a subject into a ''close'' category. Although they are not explored here, weighted kappa statistics providing such credit may alternatively be used in studies of interobserver agreement using multinomial outcomes [18].

On application of an alternative parameterization for the case of multinomial outcomes [8], the probability that each of the n raters agree (exactly) on category j can now be obtained as:

$$P\left(j_{(1)},j_{(2)},\ldots,j_{(n)}\right)=\pi_j^n\left[1+\kappa\sum_{s=1}^{n-1}Z_s+\kappa^2\sum_{s<l,s=1}^{n-1}Z_sZ_l+\ldots+\kappa^{n-1}Z_1Z_2\ldots Z_{n-1}\right]\times\left[\prod_{r=0}^{n-2}(1+r\kappa)\right]^{-1},$$

where

$$Z_s=\frac{s-\pi_j}{\pi_j},\quad \text{for}\quad s=1,2,\ldots,n-1$$

As an example, the probability mass function for the case of $k=4$ categories and $n=3$ raters is represented as:

$$\begin{aligned}
Pr_1(\pi_1,\kappa)&=P(1,1,1,1)=\pi_1\left(\pi_1^2+3\kappa\pi_1-2\kappa\pi_1^2+2\kappa^2-3\kappa^2\pi_1+\kappa^2\pi_1^2\right)(1+\kappa)^{-1}\\
Pr_2(\pi_2,\kappa)&=P(2,2,2,2)=\pi_2\left(\pi_2^2+3\kappa\pi_2-2\kappa\pi_2^2+2\kappa^2-3\kappa^2\pi_2+\kappa^2\pi_2^2\right)(1+\kappa)^{-1}\\
Pr_3(\pi_3,\kappa)&=P(3,3,3,3)=\pi_3\left(\pi_3^2+3\kappa\pi_3-2\kappa\pi_3^2+2\kappa^2-3\kappa^2\pi_3+\kappa^2\pi_3^2\right)(1+\kappa)^{-1}\\
Pr_0(\pi_1,\pi_2,\pi_3,\kappa)&=P(\text{Disagree})=1-Pr_1(\pi_1,\kappa)-Pr_2(\pi_2,\kappa)-Pr_3(\pi_3,\kappa)
\end{aligned}$$

Using this notation, we may now define the chi-square GOF statistic as:

$$X^2=\sum_{i=0}^{k}\frac{\left(m_i-N\widehat{P}r_i(\pi_1,\ldots,\pi_k,\kappa)\right)^2}{N\widehat{P}r_i(\pi_1,\ldots,\pi_k,\kappa)},\qquad(1)$$

where m_i represent the observed frequencies and $\widehat{P}r_i(\pi_1,\ldots,\pi_k,\kappa)$ is obtained by substituting the maximum likelihood estimates of π_j and κ into $Pr_i(\pi_1,\ldots,\pi_k,\kappa)$, respectively. Provided the expected cell counts are not too