The International Journal of Biostatistics

Volume 6, Issue 1

2010

Article 31

Sample Size Requirements for Interval Estimation of the Kappa Statistic for Interobserver Agreement Studies with a Binary Outcome and Multiple Raters

Allan Donner, *The University of Western Ontario*

Michael A. Rotondi, *The University of Western Ontario*

Recommended Citation:

Donner, Allan and Rotondi, Michael A. (2010) "Sample Size Requirements for Interval Estimation of the Kappa Statistic for Interobserver Agreement Studies with a Binary Outcome and Multiple Raters," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 31.

DOI: 10.2202/1557-4679.1275

Sample Size Requirements for Interval Estimation of the Kappa Statistic for Interobserver Agreement Studies with a Binary Outcome and Multiple Raters

Allan Donner and Michael A. Rotondi

Abstract

Sample size requirements that achieve a prespecified expected lower limit for a confidence interval about the intraclass kappa statistic are supplied for the case of multiple raters and a binary outcome variable. The expected lower confidence limit achievable for a given number of subjects and raters is also presented. These results should be useful in the planning stages of an interobserver agreement study in which the focus is on interval estimation rather than hypothesis-testing.

KEYWORDS: interobserver agreement, interrater agreement, reliability, kappa

Author Notes: Prof. Allan Donner is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Michael A. Rotondi is the recipient of an NSERC Canada Graduate Scholarship.

1 Introduction

It is now widely accepted that interval estimation is preferable to P -values alone for the presentation of study findings across a wide range of disciplines. This is because a properly calculated confidence interval (CI) will provide a range of values within which the investigator can have a pre-specified degree of confidence that it will contain the population parameter of interest. As stated by Gardner and Altman (2007), “confidence intervals become relevant whenever an inference is to be made from the study results to the wider world.” At the same time, it is also agreed that the strength of such inferences is greatly enhanced by a formal predetermined assessment of the required sample size.

Traditionally this has been accomplished using power considerations, although there is a growing literature providing sample size requirements from the perspective of confidence interval construction. Regardless of method, very few articles provide sample size requirements for inferences concerning the coefficient of interobserver agreement, κ , while authors who do often focus on the case of a continuous outcome measure (e.g., Donner and Eliasziw (1987), Giraudeau and Mary (2001), Walter, Eliasziw, and Donner (1998)). As stated by Shrout (1998) in the context of psychiatric research, interobserver agreement studies designed without attention to the required size of sample are often too small to provide conclusive results.

Altaye, Donner, and Klar (2001b) present sample size formulae using a power-based approach for the case of multiple raters and a binary outcome. However, the required solution is computationally intensive and only a limited number of results were presented in tabular form. Sample size requirements restricted to the case of two raters using a confidence interval approach were also presented by Donner (1999) for the case of a binary outcome.

Our objective here is to present sample size requirements for interobserver agreement studies involving two or more raters and a binary outcome from a confidence interval perspective. Since the aim of an investigator is usually to achieve a level of agreement exceeding a certain minimum value, it is our experience that interest usually focuses on the lower limit of one-sided confidence interval for κ . Thus in this paper our specific aim is to provide sample size requirements in terms of the number of subjects N and raters n that assure that the expected lower bound of a 95 % confidence limit for κ is no less than a specified threshold value κ_L . Since this approach takes into account the uncertainty inherent in the estimation procedure, it would seem preferable to the frequently adopted practice of observing whether the point estimate alone exceeds a specified threshold.

As budgetary and/or other practical restrictions often place an upper bound on the number of subjects or raters that can be enrolled in an interobserver agreement study, we also present the expected lower confidence interval bound for κ that is achievable for pre-specified values of N and n .

2 Methods

Let X_{ij} denote the binary rating for the i th subject, $i = 1, 2, \dots, N$ assigned to the j th rater, $j = 1, 2, \dots, n$ and let $\pi = P(X_{ij} = 1)$ denote the probability that the rating is a success, assumed constant across raters. An estimate of κ may be obtained by applying a standard one-way analysis of variance to the binary ratings. Following Fleiss (1981), the intraclass kappa statistic may be calculated as:

$$\hat{\kappa}_C = \frac{MSA - MSW}{MSA + (n - 1)MSW} \quad (1)$$

where MSA and MSW denote the mean square errors among and within subjects respectively.

Bahadur (1961) proposed a parsimonious model for correlated binary data that was subsequently applied by Altaye et al. (2001b) to modelling interrater agreement among multiple raters. Letting $X_i = \sum_{j=1}^n X_{ij}$ denote the total number of success for subject i , then the joint probability function for this model is given by:

$$P(X_i = x_i) = \binom{n}{x_i} \pi^{x_i} (1 - \pi)^{n - x_i} \times \left[1 + \sum_{i < j} \kappa Z_i Z_j + \sum_{i < j < k} \kappa_3 Z_i Z_j Z_k + \dots + \kappa_n Z_1 Z_2 \dots Z_n \right] \quad (2)$$

for

$$\begin{aligned} Z_i &= \frac{x_i - \pi}{(\pi(1 - \pi))^{1/2}} \\ x_i &= \sum_{j=1}^n X_{ij} \\ \kappa &= E(Z_i Z_j) \\ &\vdots \\ \kappa_n &= E(Z_1 Z_2 \dots Z_n) \end{aligned}$$

For interobserver agreement studies, the parameter of interest is the second-order correlation κ , which can be shown to be equivalent to the intraclass correlation coefficient as obtained from a one-way random effects model. For binary data, κ may be defined as:

$$\kappa = \frac{E(X_{ij}X_{ik}) - \pi^2}{\pi(1 - \pi)}, \text{ for } j \neq k \quad (3)$$

The parameters in (2) can be estimated directly from the model using the method of maximum likelihood. However, the number of parameters that need to be estimated proliferates rapidly as n increases, and thus a solution to the likelihood equation quickly becomes intractable. Therefore, Altaye et al. (2001b) obtained the maximum likelihood estimators of κ , π , $\dots$, κ_n using a reparameterization presented by George and Bowman (1995). For example, in the case $n = 3$ raters, the model (2) becomes:

$$\begin{aligned} P_0(\kappa, \kappa_3, \pi) &= P(X_i = 0) \\ &= (1 - \pi)^3 + 3\kappa\pi(1 - \pi)^2 - \kappa_3(\pi(1 - \pi))^{3/2} \\ P_1(\kappa, \kappa_3, \pi) &= P(X_i = 1) \\ &= 3\pi(1 - \pi)^2 - 3\kappa\pi(1 - \pi)(2 - 3\pi) + 3\kappa_3(\pi(1 - \pi))^{3/2} \\ P_2(\kappa, \kappa_3, \pi) &= P(X_i = 2) \\ &= 3\pi^2(1 - \pi) + 3\kappa\pi(1 - \pi)(1 - 3\pi) - 3\kappa_3(\pi(1 - \pi))^{3/2} \\ P_3(\kappa, \kappa_3, \pi) &= P(X_i = 3) \\ &= \pi^3 + 3\kappa\pi^2(1 - \pi) + \kappa_3(\pi(1 - \pi))^{3/2} \end{aligned}$$

This model may be reparameterized as:

$$\begin{aligned} P_0(\kappa, \pi) &= P(X_i = 0) = (1 - \pi)^3 + \kappa\pi((1 - \pi)^2 + (1 - \pi)) \\ P_1(\kappa, \pi) &= P(X_i = 1) = 3\pi(1 - \pi)^2(1 - \kappa) \\ P_2(\kappa, \pi) &= P(X_i = 2) = 3\pi^2(1 - \pi)(1 - \kappa) \\ P_3(\kappa, \pi) &= P(X_i = 3) = \pi^3 + \kappa\pi(1 - \pi^2) \end{aligned}$$

using the recursive definition for higher order correlations, κ_s ($s = 3, \dots, n$), as described in Altaye et al. (2001b). Additional details regarding this reparameterization are available in Altaye et al. (2001b), Altaye, Donner, and Eliasziw (2001a).

The observed ratings and the corresponding probabilities can now be summarized as shown in Table 1. Note that, when the number of raters is two, this model reduces to the common correlation model described by many authors (e.g., Bloch and Kraemer (1989)).

Table 1: Data layout for $n = 3$ raters

Category	Ratings	Frequency	Probability
0	(0,0,0)	m_0	$P_0(\kappa, \pi)$
1	(0,0,1), (0,1,0), (1,0,0)	m_1	$P_1(\kappa, \pi)$
2	(1,1,0), (1,0,1), (0,1,1)	m_2	$P_2(\kappa, \pi)$
3	(1,1,1)	m_3	$P_3(\kappa, \pi)$
Total		N	1

Let m_j , $j = 1, 2, \dots, n$ denote the number of subjects whose x_i value (the total number of successes for subject i) equals j for $j = 0, 1, \dots, n$. Using a data layout similar to Table 1, the observed frequencies $m_0, m_1, \dots, m_n$, generated in the case of n raters are seen to follow a multinomial distribution conditional on N . The probabilities $P_i(\kappa, \pi)$ can now be estimated by replacing π and κ_s ($s = 3, \dots, n$) by their maximum likelihood estimates in (2). It follows that under the original model, the Pearson χ^2 statistic,

$$X^2 = \sum_{i=0}^{l-1} \frac{(m_i - NP_i(\hat{\kappa}, \hat{\pi}))^2}{NP_i(\hat{\kappa}, \hat{\pi})} \quad (4)$$

has a limiting chi-square distribution with $l - r - 1$ degrees of freedom, where l is the number of nonredundant probability cells and r is the number of unspecified parameters in the model that need to be estimated. Note that assuming all disagreements are treated equally, inferences may be sharpened by pooling all disagreeing cells (i.e. Categories 1 and 2) into a single cell representing disagreement. Under these circumstances the X^2 statistic (4) has a limiting chi-square distribution on 1 degree of freedom, corresponding to the reduction in the number of estimated parameters under the reparameterization, and the pooling of all disagreeing cells. Assuming the required sample size is reasonably large, a one-sided confidence lower confidence limit about $\hat{\kappa}$ may now be obtained by finding the minimum admissible root, κ_L , to the equation $X^2 = \chi_{1,1-2\alpha}^2$, where $\chi_{1,1-2\alpha}^2$, is the 100(1 - α) percentile point of the chi-square distribution with one degree of freedom (Altaye et al., 2001a).

We now proceed by pre-specifying the anticipated value for κ (denoted by κ_0) in the proposed study. This value may be obtained from a pilot study or from a suitable external source. The desired value of κ_L should correspond to what the investigators consider a minimal acceptable value for the purposes at hand. In this regard guidelines provided by Landis and Koch (1977) may be helpful, where values of κ given by 0.4, 0.6, and 0.8 can be characterized as “moderate”, “substantial” and “almost perfect” agreement, respectively. Thus if the investigator wishes to be reasonably sure of achieving a “substantial” level of interobserver agreement, he may set $\kappa_L = 0.6$. An estimate of π , the anticipated prevalence rate, is also required, and may often be obtained from the literature.

After pre-specification of κ_0 , κ_L and π , an iterative procedure may be used to determine the required minimum sample size (N) that will ensure that a one-sided 95 % confidence interval for κ has an expected lower bound of κ_L . That is, an iterative loop can be used in solving equation (4) to find the minimum root of $X^2 = \chi_{1,1-2\alpha}^2 = 2.71$ for the desired values of κ_0 and π . Similar calculations can be used to provide the expected value of κ_L for a fixed number of available subjects (N) and raters. All calculations for the following tables were performed in the R environment (R Core Development Team, 2010). In addition to these values, the accompanying online supplement provides code for other values of κ_0 , κ_L , π and n .

3 Results

Table 2 provides the number of subjects required (N) to ensure that the lower one-sided 95 % confidence limit for κ is no less than an expected lower bound of κ_L for various values of κ_0 , n and π .

For example suppose an investigator wishes to assure with 95 % confidence that a “substantial” level of interobserver agreement is achieved when the anticipated value of $\kappa_0 = 0.8$. Then assuming $\pi = 0.1$, the number of subjects required to achieve a value of κ_L no less than 0.6 decreases from 116 to 62 as the number of raters increases from 2 to 4. On the other hand if the investigator feels that an anticipated outcome prevalence of $\pi = 0.3$ is reasonable, the required number of subjects is markedly reduced, from 52 ($n = 2$) to 25 ($n = 4$) respectively. However, if an anticipated value for κ_0 of 0.8 is considered too optimistic, and the true level of interobserver agreement is instead assumed to be $\kappa_0 = 0.6$, the number of subjects required to ensure “moderate” agreement under these circumstances now ranges from 140 to 76

Table 2: Number of subjects N required to ensure that the expected lower limit of a 95 % one-sided confidence limit for κ is no less than κ_L

κ_0	κ_L	π	Number of Raters (n)			
			2	3	4	5
0.50	0.40	0.10	559	373	301	255
		0.30	264	146	112	95
		0.50	228	120	89	76
0.60	0.40	0.10	140	94	76	64
		0.30	66	37	28	24
		0.50	57	30	23	19
0.70	0.60	0.10	463	311	247	207
		0.30	205	124	99	87
		0.50	174	102	81	73
0.80	0.60	0.10	116	78	62	52
		0.30	52	31	25	22
		0.50	44	26	21	19

as n increases from 2 to 4 raters (for $\pi = 0.1$). It is also clear that as π approaches 0.5 the required value of N becomes substantially less, particularly as π increases from 0.1 to 0.3. Thus a conservative strategy in practice would be to take a reasonable lower bound for the value of π . (Since the resulting value of N is symmetric about $\pi = 0.5$, taking a reasonable upper bound is conservative when $\pi > 0.5$). Finally, in terms of the number of raters, it is clear that the greatest reductions in N are achieved when the number of raters is increased from two to three raters, with gradually diminishing returns after that point.

In many studies of interobserver agreement there will be a well-defined practical limit on the number of subjects and/or raters that can be enrolled. In this case Table 3 may be used to assess the expected value of the 95 % lower confidence limit that can be achieved for given values of κ_0 and π . For example, if the maximum number of subjects that can be enrolled is $N = 100$, and $\kappa_0 = 0.7$, then in order to be 95 % confident of achieving a “substantial” level of interobserver agreement, π must be at least 0.3 and the study must recruit at least four raters. However if only a “moderate” level of interobserver agreement is required then two raters will be sufficient for values of π as low as 0.1.

Many studies will be restricted to enrolling at most 50 subjects. In this case it is virtually impossible to achieve a value of $\kappa_L = 0.6$ unless the anticipated value of the coefficient of interobserver agreement is at least 0.8.

4 Discussion

As statistical methods for the analysis of data arising from interobserver agreement studies continue to develop, it becomes increasingly important to consider the design features that drive these analyses. These considerations include the need to recruit a suitable number of both subjects (N) and raters (n). In this paper we have presented the values of N and n that assure with 95 % confidence that the coefficient of interobserver agreement is no less than a specified threshold value. Although the choice of this value is inevitably somewhat arbitrary, we have suggested that the guidelines provided by Landis and Koch (1977) may be useful.

Further note that the anticipated value for κ_0 may drastically differ from that which is observed upon completion of the experiment. This may in practice limit the experimenters ability to achieve the desired lower bound for κ . For this reason, a detailed sensitivity analysis is advised to ensure that minor variations in π and κ_0 do not produce extreme changes in the required sample size. Nonetheless, the CI approach provides additional insight into the anticipated precision about κ that can reasonably be achieved in the planned study.

The goodness-of-fit procedure has been shown to provide confidence interval coverage close to nominal for studies that recruit approximately 25 subjects or more (e.g. Altaye et al. (2001b), Donner and Eliasziw (1992)). For studies smaller than this, the procedure gradually becomes less stable. However such studies, in any event, will yield very wide confidence intervals owing to their low precision.

In order to avoid undue complexity the results given in Tables 2 and 3 are derived under the assumption that there is no rater bias, that is, that each rater may be characterized by the same underlying success rate π . As discussed by Landis and Koch (1977) this assumption is most appropriate when the emphasis is directed at the reliability of the measurement process itself.

Table 3: Expected lower limit of a 95 % one-sided confidence limit for κ for a fixed available sample size, N_{max}

N_{max}	κ_0	π	Number of Raters (n)			
			2	3	4	5
25	0.50	0.10	0.09	0.16	0.18	0.19
		0.30	0.16	0.26	0.29	0.31
		0.50	0.17	0.28	0.31	0.33
25	0.60	0.10	0.16	0.23	0.26	0.28
		0.30	0.26	0.36	0.39	0.40
		0.50	0.28	0.38	0.41	0.42
25	0.70	0.10	0.24	0.32	0.35	0.38
		0.30	0.37	0.46	0.49	0.51
		0.50	0.40	0.49	0.51	0.52
25	0.80	0.10	0.32	0.41	0.46	0.49
		0.30	0.49	0.57	0.60	0.62
		0.50	0.52	0.60	0.62	0.63
50	0.50	0.10	0.18	0.24	0.26	0.27
		0.30	0.26	0.33	0.35	0.36
		0.50	0.28	0.34	0.37	0.38
50	0.60	0.10	0.27	0.33	0.35	0.37
		0.30	0.37	0.43	0.45	0.46
		0.50	0.38	0.45	0.47	0.47
50	0.70	0.10	0.36	0.43	0.46	0.48
		0.30	0.48	0.54	0.56	0.57
		0.50	0.50	0.55	0.57	0.58
50	0.80	0.10	0.47	0.54	0.57	0.60
		0.30	0.60	0.65	0.67	0.68
		0.50	0.62	0.66	0.68	0.69
100	0.50	0.10	0.26	0.31	0.33	0.34
		0.30	0.33	0.38	0.39	0.40
		0.50	0.34	0.39	0.40	0.41
100	0.60	0.10	0.36	0.41	0.43	0.44
		0.30	0.44	0.48	0.49	0.50
		0.50	0.45	0.49	0.51	0.51
100	0.70	0.10	0.47	0.51	0.53	0.55
		0.30	0.55	0.59	0.60	0.61
		0.50	0.56	0.60	0.61	0.61
100	0.80	0.10	0.58	0.63	0.65	0.66
		0.30	0.67	0.70	0.71	0.72
		0.50	0.68	0.71	0.72	0.72

Table 3 (continued)

N_{max}	κ_0	π	Number of Raters (n)			
			2	3	4	5
150	0.50	0.10	0.31	0.34	0.36	0.37
		0.30	0.36	0.40	0.41	0.42
		0.50	0.37	0.41	0.42	0.43
150	0.60	0.10	0.41	0.44	0.46	0.47
		0.30	0.47	0.50	0.51	0.52
		0.50	0.48	0.51	0.52	0.53
150	0.70	0.10	0.51	0.55	0.57	0.58
		0.30	0.58	0.61	0.62	0.62
		0.50	0.59	0.62	0.63	0.63
150	0.80	0.10	0.63	0.66	0.68	0.69
		0.30	0.69	0.72	0.73	0.73
		0.50	0.70	0.73	0.73	0.74
200	0.50	0.10	0.33	0.36	0.38	0.39
		0.30	0.38	0.41	0.42	0.43
		0.50	0.39	0.42	0.43	0.44
200	0.60	0.10	0.43	0.46	0.48	0.49
		0.30	0.49	0.52	0.53	0.53
		0.50	0.50	0.52	0.53	0.54
200	0.70	0.10	0.54	0.57	0.59	0.60
		0.30	0.60	0.62	0.63	0.63
		0.50	0.61	0.63	0.64	0.64
200	0.80	0.10	0.65	0.68	0.70	0.71
		0.30	0.71	0.73	0.74	0.74
		0.50	0.72	0.74	0.74	0.75

Given the approximate nature of sample size estimation, the assumption of no rater bias should generally be sufficient for sample size planning. Thus, if the differences in rater prevalence are small, their effects will be averaged out in both the estimation of sample size and in the subsequent statistical analysis. It could also be argued that if these biases are substantial, it would be misleading to go further in attempting to measure other aspects of inter-observer agreement (see Zwick (1988)). In the analysis phase of the study, a test for no rater bias is provided by the well-known Cochran's Q-statistic, as illustrated by Fleiss (1981).

The approach presented here does not take into account the costs involved in recruiting subjects and/or raters, as has been done for the case of a continuous outcome. The overall cost would typically depend on C_1 , the cost of recruiting a subject, C_2 , the cost of recruiting a rater and C_3 , the cost of obtaining a measurement on a subject that has already been recruited. This leads naturally to an overall linear cost model of the form $C = NC_1 + nC_2 + NnC_3$. One could then determine the optimal N and n that yield a specified value of κ_L while minimizing the overall cost C . A similar approach was taken by Eliasziw and Donner (1987) for the case of a continuous outcome variable using a power-based approach. This work is the subject of future research.

References

- Altaye, M., A. Donner, and M. Eliasziw (2001a): "A general goodness-of-fit approach for inference procedures concerning the kappa statistic," *Stat. Med.*, 20, 2479–2488.
- Altaye, M., A. Donner, and N. Klar (2001b): "Inference procedures for assessing interobserver agreement among multiple raters," *Biometrics.*, 57, 584–588.
- Bahadur, R. (1961): "A representation of the joint distribution of responses to n dichotomous items," in H. Solomonon, ed., *Studies in Item Analysis and Prediction*, Stanford, California: Stanford Mathematical Studies in the Social sciences.
- Bloch, D. and H. C. Kraemer (1989): "2x2 kappa coefficients: Measure of agreement or association," *Biometrics.*, 45, 269–287.
- Donner, A. (1999): "Sample size requirements for interval estimation of the intraclass kappa statistic," *Com. Stat. - Sim. Comp.*, 28, 415429.
- Donner, A. and M. Eliasziw (1987): "Sample size requirements for reliability studies," *Stat. Med.*, 6, 441–448.
- Donner, A. and M. Eliasziw (1992): "A goodness-of-fit approach to inference procedures for the kappa statistics: Confidence interval construction, significant testing and sample size estimation," *Stat. Med.*, 11, 1511–1519.
- Eliasziw, M. and A. Donner (1987): "A cost-function approach to the design of reliability studies," *Stat. Med.*, 6, 647–655.
- Fleiss, J. (1981): *Statistical Methods for Rates and Proportions*, Wiley, New York.
- Gardner, M. J. and D. G. Altman (2007): "Estimating with confidence," in D. G. Altman et al., eds., *Statistics with Confidence*, British Medical Journal Books.

- George, E. O. and D. Bowman (1995): "A full likelihood procedure for analyzing exchangeable binary data," *Biometrics.*, 51, 512–523.
- Giraudeau, B. and J. Y. Mary (2001): "Planning a reproducibility study: How many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient," *Stat. Med.*, 21, 3205–3214.
- Landis, J. R. and G. G. Koch (1977): "The measurement of observer agreement for categorical data," *Biometrics.*, 33, 159–179.
- R Core Development Team (2010): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Shrout, P. E. (1998): "Measurement reliability and agreement in psychiatry," *Stat. Meth. Med. Res.*, 7, 301–317.
- Walter, S. D., M. Eliasziw, and A. Donner (1998): "Sample size and optimal designs for reliability studies," *Stat. Med.*, 17, 101–110.
- Zwick, R. (1988): "Another look at interrater agreement," *Psych. Bull.*, 103, 374–378.