

Interval Estimation of the Kappa Coefficient with Binary Classification and an Equal Marginal Probability Model

Jun-Mo Nam

Biostatistics Branch, National Cancer Institute
Executive Plaza South, Room 8028, 6120 Executive Boulevard, MSC 7368,
Rockville, Maryland 20892-7368, U.S.A.
email: namj@mail.nih.gov

SUMMARY. We derive a likelihood score method for interval estimation of the intraclass version of the kappa coefficient of agreement with binary classification using a general theory of Bartlett (1953, *Biometrika* 40, 306-317). By exact evaluation, we investigate statistical properties of the score method, the chi-square goodness-of-fit procedure (Donner and Eliasziw, 1992, *Statistics in Medicine* 11, 1511-1519; Hale and Fleiss, 1993, *Biometrics* 49, 523-534), and a crude confidence interval for small and medium sample sizes. Actual coverage percentages of the score and chi-square methods are satisfactorily close to the nominal confidence coefficient, while that of the crude method is quite unsatisfactory. The expected length of the score method is shorter than that of the chi-square procedure when the response rate is very small or very large.

KEY WORDS: Expected coverage probability; Expected length of interval; Interval estimation; Intraclass correlation; Kappa coefficient of agreement; Score method.

1. Introduction

There are two basic underlying models for assessing the degree of agreement on a binary scale. The first model assumes the marginal probabilities of positive classification with two raters are different, and Cohen's (1960) kappa is based on this model. The second model assumes the same marginal positive probability for each rater and leads to the intraclass version of the kappa, which is identical to Scott's (1955) index. In this article, we will limit our attention to evaluation of agreement under the second model, i.e., the reliability of a single rater based on rating a sample of subjects twice (Hale and Fleiss, 1993). Recently, Donner and Eliasziw (1992) and Hale and Fleiss (1993) have independently developed the estimation method using a chi-square goodness-of-fit statistic and Cornfield's test-based method, respectively. In this article, we derive interval estimation that possesses asymptotically optimal statistical properties and compare it with other methods for small and medium sample sizes by exact evaluation.

2. Model and Notation

Suppose that two comparable raters each rate a sample of n subjects independently, with ratings denoted by either positive (+) or negative (-) responses, or a single rater rates a set of n samples twice with the second evaluation done without knowledge of the first one. The pairs of ratings are classified in three categories. With subscripts presenting the number of positive ratings in a pair, let x_2 , x_1 , and x_0 denote the observed numbers of n pairs ratings and P_2 , P_1 , and P_0 the corresponding probabilities. Let p denote the probability of a positive response and $q = 1 - p$, i.e., $\Pr(+)=p$ and $\Pr(-)=q$.

The kappa coefficient, κ , is a measurement of the correlation between two ratings in a pair (interrater or intrarater reliability). The intraclass correlation in a pair may be defined as $\rho = (P_2 - p^2)/(pq) = (P_0 - q^2)/(pq)$, so $P_2 = p^2 + pq\rho$, $P_0 = q^2 + pq\rho$, and $P_1 = 1 - P_2 - P_0 = 2pq(1 - \rho)$. The actual probability of agreement is $p_o = P_2 + P_0$ and the probability of agreement by chance is $p_e = p^2 + q^2$, where $p = P_2 + P_1/2$. Since the standard definition of kappa is $\kappa \equiv (p_o - p_e)/(1 - p_e) = \rho$, kappa is the same as the intraclass correlation.

3. Interval Estimation of Coefficient of Agreement

Consider the distribution of $\mathbf{x}' = (x_2, x_1, x_0)$ under the multinomial model in Section 2. The log likelihood is expressed as

$$\ln L = x_2 \ln\{p(p + q\kappa)\} + x_1 \ln\{2pq(1 - \kappa)\} + x_0 \ln\{q(q + p\kappa)\},$$

where $q = 1 - p$. Kappa, κ , is the parameter of interest and p is a nuisance parameter.

The maximum likelihood estimators (MLEs) of κ and p are $\hat{\kappa} = (4x_0x_2 - x_1^2)/\{(2x_0 + x_1)(2x_2 + x_1)\}$ and $\hat{p} = (2x_2 + x_1)/(2n)$, respectively. The asymptotic variance of $\hat{\kappa}$ is

$$\text{var}(\hat{\kappa}) = (1 - \kappa)\{(1 - \kappa)(1 - 2\kappa) + \kappa(2 - \kappa)/(2pq)\}/n \quad (1)$$

(cf., Hale and Fleiss, 1993). The estimated variance of $\hat{\kappa}$ is obtained by replacing κ and p by $\hat{\kappa}$ and \hat{p} in (1). The $100 \times (1 - \alpha)$ percent crude confidence limits of κ are found by

$$\hat{\kappa} \pm z_{(1-\alpha/2)} \{\text{var}(\hat{\kappa})\}_{\kappa=\hat{\kappa}, p=\hat{p}}^{1/2} \quad (2)$$

where $z_{(1-\alpha/2)}$ is the $100 \times (1 - \alpha/2)$ percentile point of the

standardized normal distribution. Donner and Eliasziw (1992) proposed a $100\% \times (1 - \alpha)$ confidence interval for κ based on a goodness-of-fit statistic,

$$X_G^2 = \sum_{i=0}^2 \{x_i - nP_i(\kappa, \hat{p})\}^2 / \{nP_i(\kappa, \hat{p})\} = z_{\alpha/2}^2, \quad (3)$$

where $\hat{p} = (2x_2 + x_1)/(2n)$, $P_2(\kappa, \hat{p}) = \hat{p}^2 + \hat{p}\hat{q}\kappa$, $P_1(\kappa, \hat{p}) = 2\hat{p}\hat{q}(1 - \kappa)$, and $P_0(\kappa, \hat{p}) = \hat{q}^2 + \hat{p}\hat{q}\kappa$, with $\hat{q} = 1 - \hat{p}$. The confidence limits are two admissible roots of a cubic equation of κ . Hale and Fleiss (1993) presented a Cornfield-type method. The limits of the interval with a $(1 - \alpha)$ coefficient are obtained by solving the following equation:

$$z^2 = \{x_2 - nP_2(\kappa, \hat{p})\}^2 \times \left\{ \frac{1}{nP_0(\kappa, \hat{p})} + \frac{4}{nP_1(\kappa, \hat{p})} + \frac{1}{nP_2(\kappa, \hat{p})} \right\} = z_{(\alpha/2)}^2.$$

The goodness-of-fit and Cornfield's test-based methods are, in fact, algebraically the same and will henceforth be referred as the DE & HF method in this article.

Denote the first-order derivatives of the log likelihood as

$$S_\kappa(\kappa, p) \equiv \frac{\partial \ln L}{\partial \kappa} = \frac{1}{1 - \kappa} \left(\frac{x_2}{p + q\kappa} + \frac{x_0}{q + p\kappa} - n \right)$$

$$S_p(\kappa, p) \equiv \frac{\partial \ln L}{\partial p} = \frac{x_1 + x_2 - (n + x_1)p}{pq} + \frac{\{x_2 - x_0\kappa - (x_0 + x_2)(1 - \kappa)p\}(1 - \kappa)}{(p + q\kappa)(q + p\kappa)}, \quad (4)$$

where $q = 1 - p$. Using results of Bartlett (1953), the asymptotic variance of the score evaluated at $p = \hat{p}$, where \hat{p} is the MLE of p for a given value of κ , is obtained by

$$\text{var}\{S_\kappa(\kappa, \hat{p})\} = 2npq / \{2pq(1 - \kappa)(1 - 2\kappa) + \kappa(2 - \kappa)\}(1 - \kappa).$$

Note that Bartlett's formulation is essentially similar to that of Cox and Hinkley (1974). The approximate $1 - \alpha$ confidence limits are two solution to the equation

$$z_s^2(\kappa, \hat{p}) = \{S_\kappa(\kappa, \hat{p})\}^2 / \text{var}\{S_\kappa(\kappa, \hat{p})\}_{p=\hat{p}} = z_{(\alpha/2)}^2$$

or

$$\left(\frac{x_2}{\hat{p} + \hat{q}\kappa} + \frac{x_0}{\hat{q} + \hat{p}\kappa} - n \right)^2 \left\{ \frac{2\hat{p}\hat{q}(1 - \kappa)(1 - 2\kappa) + \kappa(2 - \kappa)}{2n\hat{p}\hat{q}(1 - \kappa)} \right\} = z_{(\alpha/2)}^2. \quad (5)$$

The MLE, \hat{p} , is the solution of a cubic equation, $S_p(\kappa, \hat{p}) = 0$, i.e.,

$$a_0\hat{p}^3 + a_1\hat{p}^2 + a_2\hat{p} + a_3 = 0, \quad (6)$$

where $a_0 = 2n(1 - \kappa)^2$, $a_1 = -\{3n(1 - \kappa) + x_2 - x_0\}(1 - \kappa)$, $a_2 = 2x_2 + x_1 - 2(2n - x_0)\kappa + n\kappa^2$, and $a_3 = (x_1 + x_2)\kappa$. Only one root (see Appendix) is appropriate as the MLE of p and the other two roots (one is negative and the other is greater than one) are irrelevant. The confidence limits are two solutions that satisfy equation (5) with (A.1). They may be found iteratively using Newton-Raphson or the secant method.

4. Exact Evaluation

We evaluate exact confidence coefficient for the crude, DE & HF, and score methods by numerical evaluation of multinomial probabilities at all possible sampling points. The method is similar to the exact evaluation of interval estimation of the ratio of two binomial parameters (Gart and Nam, 1988). The exact probabilities of the interval estimation covering κ is

$$P = \sum_{x \in R} \text{Pr}(x | n, \kappa, p),$$

where R is the region of sampling points in which the confidence interval contains κ . The $\text{Pr}(x | n, \kappa, p)$ is the multinomial distribution of x defined in Section 2. We calculate exact confidence coefficients for the crude, DE & HF, and score methods for a wide range of p and κ for small and medium sample sizes. Table 1 provides typical values of actual coverage probabilities of the methods corresponding to a nominal 95% confidence coefficient for $n = 20$ and 40. The actual coefficient of the crude method is noticeably smaller than the nominal value, particularly for $n = 20$. The DE & HF method and the score method yield actual coefficients satisfactorily

Table 1
Actual coverage percentage for nominal 95% confidence interval for κ

p	κ	n = 20			n = 40		
		Crude	DE & HF	Score	Crude	DE & HF	Score
0.1	0.1	30.4	96.7	93.5	51.8	96.4	96.4
	0.3	48.0	97.6	95.1	73.0	96.7	95.9
	0.5	60.5	95.7	97.0	81.6	93.9	96.0
	0.7	58.1	92.0	96.8	82.0	92.8	95.3
	0.9	35.5	92.0	92.0	51.3	92.6	94.9
0.3	0.1	85.2	95.3	95.3	92.1	94.9	95.3
	0.3	88.6	94.9	94.9	92.9	94.4	94.8
	0.5	90.0	94.4	94.5	92.6	94.8	95.0
	0.7	89.1	95.1	95.2	91.1	94.4	95.3
	0.9	57.3	93.9	93.9	81.7	95.9	95.9

Material may be protected by copyright law. Title 17, U.S. Code

Table 2
Expected length of nominal 95% confidence interval

p	κ	n = 20		n = 40	
		DE & HF	Score	DE & HF	Score
0.1	0.1	0.725	0.513	0.573	0.446
	0.3	0.782	0.656	0.639	0.588
	0.5	0.817	0.732	0.661	0.636
	0.7	0.829	0.761	0.634	0.607
	0.9	0.813	0.751	0.543	0.504
0.3	0.1	0.714	0.706	0.560	0.579
	0.3	0.736	0.745	0.572	0.582
	0.5	0.714	0.718	0.544	0.546
	0.7	0.643	0.637	0.472	0.470
	0.9	0.503	0.488	0.332	0.327

close to the nominal one. The discrepancies between the two methods in the limit may be caused by differences in tail probabilities and also by differences in the efficiency of estimation. Denoting $\hat{\kappa}_l$ and $\hat{\kappa}_u$ as lower and upper limits of an interval, we calculate the expected length of each of two methods by

$$E(\hat{l}) = \sum_x \hat{l} \Pr(x | n, \kappa, p),$$

where $\hat{l} = \hat{\kappa}_u - \hat{\kappa}_l$ and summation is over all possible sampling points. Table 2 shows that the expected length is inversely related to sample size. The score method yields a considerably shorter expected interval length than the DE & HF method when $p = 0.1$, while the two methods are hardly different when $p = 0.3$. The tables for $p = 0.7$ and 0.9 are the same as those in Tables 1 and 2 for $p = 0.3$ and 0.1 , respectively, i.e., they are symmetric with respect to $p = 0.5$.

5. An Example

Twenty pairs of male siblings from black American families were examined for HIV seropositivity. Two, one, and 17 pairs were classified as both brothers in a pair being positive, only one being positive, and both being negative, respectively. We want to assess the intraclass correlation in pairs of siblings. The point estimate of kappa and its standard error are $\hat{\kappa} = 0.7714$ and $SE(\hat{\kappa}) = 0.2193$. From (2), (3), and (5) with (A.1), the 95% confidence interval for κ by the crude, DE & HF, and score methods are (0.3416, 1.2013), (0.2073, 0.9591), and (0.2463, 0.9620), respectively. The upper bound of the crude method is larger than one, and it is beyond the admissible range of κ . Since the point estimate of agreement between two siblings is high and the sample size is small, researchers may be more concerned with the lower limit of κ . The lower limit for the two-sided 95% confidence interval by the score method is 19% higher than that by the DE & HF method.

6. Remarks

Donner and Eliasziw (1992) have advised the use of the DE & HF method when the expected frequencies are not less than one. We find the DE & HF method and the score method give actual coverage frequencies relatively close to the nominal coefficient even when the minimum of the expected number of observations is as small as 0.4. Both the DE & HF and the score methods can always provide admissible limits, while the crude method may not.

RÉSUMÉ

Nous dérivons une méthode du score de vraisemblance pour l'estimation par intervalle de la version intraclasse du coefficient Kappa de concordance avec une classification binaire en utilisant une théorie générale de Bartlett (1953, *Biometrika* 40, 306-317). Par une évaluation exacte, nous étudions les propriétés statistiques de la méthode du score, la procédure du Chi2 d'ajustement (Donner and Eliasziw, 1992, *Statistics in Medicine* 11, 1511-1519; Hale and Fleiss, 1993, *Biometrics* 49, 523-534) et un intervalle de confiance approché pour les échantillons de petite et moyenne taille. Les pourcentages de couverture actuels des méthodes du score et du chi2 sont raisonnablement proches de l'intervalle de confiance nominal alors que la méthode approchée est n'est pas vraiment satisfaisante. La longueur attendue de l'intervalle par la méthode du score est plus courte que celle du Chi2 quand le taux de réponse est très petit ou très grand.

REFERENCES

Bartlett, M. S. (1953). Approximate confidence interval. II. More than one unknown parameter. *Biometrika* 40, 306-317.
 Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37-46.
 Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. New York: Wiley.
 Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* 11, 1511-1519.
 Gart, J. J. and Nam, J. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and correction for skewness. *Biometrics* 44, 323-338.
 Hale, C. A. and Fleiss, J. L. (1993). Interval estimation under two study designs for kappa with binary classifications. *Biometrics* 49, 523-534.
 Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* 19, 321-325.

Received February 1999. Revised August 1999.
 Accepted September 1999.

APPENDIX

The MLE of the Nuisance Parameter for a Given Value of Kappa

From (4) and (6), the MLE of p for a given value of κ , \tilde{p} , is obtained by solving the cubic equation $a_0\tilde{p}^3 + a_1\tilde{p}^2 + a_2\tilde{p} + a_3 = 0$, where $a_0 = 2n(1 - \kappa)^2$, $a_1 = -\{3n(1 - \kappa) + x_2 - x_0\}(1 - \kappa)$, $a_2 = 2x_2 + x_1 - 2(2n - x_0)\kappa + n\kappa^2$, and $a_3 = (x_1 + x_2)\kappa$. Denoting $b_i = a_i/a_0$, $i = 1, 2$ and 3 , $c_1 = b_2 - b_1^2/3$, and $c_2 = b_3 - b_1b_2/3 + 2(b_1/3)^3$, the three roots of the above cubic equation are found with a trigonometric solution. Only one root is admissible as the MLE of p , i.e.,

$$\tilde{p} = -2(-c_1/3)^{1/2} \cos(\pi/3 + \theta/3) - b_1/3, \quad (A.1)$$

where $\cos \theta = (27)^{1/2}c_2/\{2c_1(-c_1)^{1/2}\}$.

Material may be protected by copyright law (Title 17, U.S. Code)