Journal of Statistical Computation & Simulation

Vol. 00, No. 00, Month 200x, 1–15

A bootstrap method for comparing correlated kappa coefficients

S. Vanbelle† * and A. Albert†

†Medical Informatics and Biostatistics, School of Public Health, University of Liège, CHU Sart Tilman, 4000 Liège, Belgium

(Received 00 Month 200x; In final form 00 Month 200x)

Cohen's kappa coefficient is traditionally used to quantify the degree of agreement between two raters on a nominal scale. Correlated kappas occur in many settings (e.g. repeated agreement by raters on the same individuals, concordance between diagnostic tests and a gold standard) and often need to be compared. While different techniques are now available to model correlated κ coefficients, they are generally not easy to implement in practice. The present paper describes a simple alternative method based on the bootstrap for comparing correlated kappa coefficients. The method is illustrated by examples and its type I error studied using simulations. The method is also compared to the generalized estimating equations of second order and the weighted least-squares methods.

Keywords: Cohen's kappa, comparison, bootstrap

Medical Informatics and Biostatistics, School of Public Health, University of Liège, CHU Sart Tilman, 4000 Liège, Belgium, sophie.vanbelle@ulg.ac.be

1 Introduction

The kappa (κ) coefficient proposed by Cohen [1] in 1960 is widely used to assess the degree of agreement between two raters on a binary or nominal scale. It corrects the observed percentage of agreements between the raters for the effect of chance. Thus, a value of 0 implies no agreement beyond chance, whereas a value of 1 corresponds to a perfect agreement between the two raters. Correlated kappas can occur in many ways. For example, two raters may assess the same individuals at various occasions or in different experimental conditions and it may be of interest to test for homogeneity of the kappas. Alternatively, each member of a group of raters may be compared to an expert in assessing the same items on a nominal scale. Are there differences between the individual kappas obtained? The same problem arises when comparing several diagnostic tests on a binary scale (negative/positive) with respect to a gold standard. Fleiss [2] developed a method based on the chi-square decomposition for comparing two or more κ coefficients but only applicable to independent samples. McKenzie et al. [3] proposed an approach based on resampling for the comparison of two correlated κ coefficients. With the advent of generalized linear mixed models, it is now possible to model the coefficient κ as a function of covariates. Williamson et al. [4] used the generalized estimating equations of second order (GEE2) to model correlated kappas. Lipsitz et al. [5] proposed an empirical method to model independent κ coefficients. Finally, Barnhart and Williamson [6] used the weighted least-squares approach (WLS) to model correlated κ coefficients with respect to categorical covariates. All modeling techniques represent a considerable progress but they require adequate model specifications and expert programming skills. Currently, no simple method can be found in the literature for comparing several correlated κ coefficients. The present paper describes a practical and feasible alternative to the modeling techniques by expanding the resampling method based on bootstrap proposed by McKenzie et al. [3]. The original method is exposed in Section 2 and the extension detailed in Section 3. Simulations of the type I error are given in Section 4 for different levels of the kappa coefficient and different sample sizes. Results are compared to those obtained by the GEE2 and the WLS methods. The bootstrap, GEE2 and WLS methods were applied to two examples in Section 5. Finally, results are discussed in Section 6.

2 Bootstrapping two correlated kappas

Suppose that two raters classify n subjects on a binary or nominal scale at two different occasions or in two different experimental settings. Let $\hat{\kappa}_1$ and $\hat{\kappa}_2$ be the kappa coefficients obtained. Since the two agreements are assessed on the same subjects, $\hat{\kappa}_1$ and $\hat{\kappa}_2$ are correlated. Are they statistically different? Let $H_0 : \kappa_1 = \kappa_2$, the null hypothesis to be tested. The bootstrap method consists in drawing q samples (1000 is generally sufficient [3]) of size n with replacement. For each generated sample, the κ coefficient between the 2 raters is estimated in the two settings and their difference $\widehat{\kappa}_d = \widehat{\kappa}_2 - \widehat{\kappa}_1$ calculated. McKenzie et al. [3] suggested to determine the bootstrap two-sided $(1-\alpha)$-confidence interval for the $\hat{\kappa}_d$ differences, whence rejecting the null hypothesis if the confidence interval did not include 0. This approach is equivalent to using a Student's t-test and to reject H_0 at the α-significance level if

$$|t_{obs}| = \left| \frac{\overline{\kappa}_d}{SE(\kappa_d)} \right| \geq Q_t(1-\alpha/2; q-1) \quad (1)$$

where $\overline{\kappa}_d$ and $SE(\kappa_d)$ are respectively the mean and standard deviation of the q bootstrapped kappa differences and $Q_t(1-\alpha/2; q-1)$ is the upper $\alpha/2$-percentile of the Student's t distribution on $q-1$ degrees of freedom. Otherwise, H_0 is not rejected.

3 Extension to several correlated kappas

Suppose we want to compare $G \geq 2$ correlated kappa coefficients $(\kappa_1, \cdots, \kappa_G)$ i.e., to test the null hypothesis $H_0 : \ \kappa_1 = \cdots = \kappa_G$ against the alternative hypothesis $H_1 : \exists k \neq l \in \{1, \cdots, G\} : \kappa_k \neq \kappa_l$. As before, the bootstrap method will consist in drawing q samples of size n with replacement from the original data. Then, for each bootstrapped sample $(j = 1, \cdots, q)$, let $\widehat{\boldsymbol{\kappa}}_j = (\widehat{\kappa}_{1(j)}, \cdots, \widehat{\kappa}_{G(j)})'$ be the vector of the G kappa coefficients obtained. The null and alternative hypotheses can be rewritten in matrix form as follows: $H_0 : \boldsymbol{C\kappa} = \mathbf{0}$ versus $H_1 : \boldsymbol{C\kappa} \neq \mathbf{0}$, where $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_G)'$ and $\boldsymbol{C}$ the $(G-1) \times G$ patterned matrix

$$\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix}$$

Then, the test statistic is

$$T^2 = (\boldsymbol{C\overline{\kappa}})'(\boldsymbol{CSC}')^{-1}\boldsymbol{C\overline{\kappa}} \tag{2}$$

distributed as Hotelling's T^2, where $\overline{\boldsymbol{\kappa}}$ and $\boldsymbol{S}$ are respectively the sample mean vector and covariance matrix of the q bootstrapped vectors $\hat{\boldsymbol{\kappa}}$. The null hypothesis will be rejected at the α-level if

$$T^2 \geq \frac{(q-1)(G-1)}{(q-G+1)} Q_F(1-\alpha; G-1, q-G+1) \tag{3}$$

where $Q_F(1-\alpha; G-1, q-G+1)$ is the upper α-percentile of the F distribution on $G-1$ and $q-G+1$ degrees of freedom. Otherwise, H_0 will not be rejected. Note that, since "$q - G + 1$" will be large in general, the left-hand side of equation 3 can be approximated by $Q_{\chi^2}(1-\alpha; G-1)$, the $(1-\alpha)$th percentile of the chi-square distribution on $G-1$ degrees of freedom. If $\boldsymbol{c}_g$ denotes the g-th row of matrix $\boldsymbol{C}$, simultaneous confidence intervals for individual contrasts $\boldsymbol{c}_g'\boldsymbol{\kappa}$ $(g = 1, \cdots, G-1)$ given by

$$\boldsymbol{c}_g'\overline{\boldsymbol{\kappa}} \pm \sqrt{\frac{(q-1)(G-1)}{(q-G+1)} Q_F(1-\alpha; G-1, q-G+1)} \sqrt{\boldsymbol{c}_g'\boldsymbol{S}\boldsymbol{c}_g} \tag{4}$$

can be used for multiple comparison purposes.

4 Simulations

The method described in Section 3 was applied to simulated data sets in order to study the behavior of the type I error (α) of the homogeneity test for $G = 3$. Each simulation consisted in applying the bootstrap method to 3000 data sets generated under the null hypothesis $H_0 : \kappa_1 = \kappa_2 = \kappa_3$ and to determine the number of times H_0 was rejected. The simulated data set was based on 4 binary random variables X, Y, Z and V. The agreement between X and Y (κ_{XY}), X and Z (κ_{XZ}) and X and V (κ_{XV}) were compared using the bootstrap method with $q = 2000$ iterations. Simulations were repeated for 3 sample sizes (50, 75 and 100) and 5 levels of agreement (κ=0, 0.2, 0.4, 0.6 and 0.8). To obtain a given level of agreement (κ), 2 vectors of size n from binary random variables (U and W) were generated. Then, a vector of size n

with uniform random numbers between 0 and 1 was generated. Each time the random uniform number was less than or equal to the given level of agreement (κ), the value of W was changed into the value of U, otherwise it remained unchanged. The kappa coefficient was derived from the 2×2 table obtained by cross-classifying the vectors U and W. The codes for the simulations were written in R language using uniform random number generator with seed equal to 2. The method of generalized estimating equations of second order (GEE2) [4] and the weighted least square approach (WLS) [6] were also applied to the 3000 simulated data sets. Results are summarized in Table 1. It is seen that type I error rates obtained with the bootstrap method are slightly but systematically higher than the expected 5% nominal level. While the GEE2 approach appears to be optimal, the bootstrap was better than the WLS, at least for elevated κ values. However, the bootstrap method may be preferred to the GEE2 approach because of the ease of implementation in all settings as compared to the GEE2 method, which requires the writing of a lengthy and specific program for each particular problem.

5 Examples

5.1 *Deep venous thrombosis*

A study was conducted on 107 patients in the medical imaging department of the university hospital (unpublished data) to compare deep venous thrombosis

Table 1. Type I error for the comparison of $G = 3$ correlated kappa coefficients, according to κ level and sample size (figures are based on 3000 simulations each)

		κ level				
Sample size	Method	0	0.2	0.4	0.6	0.8
50	Bootstrap[a]	0.065	0.069	0.061	0.076	0.056
	GEE2	0.067	0.061	0.063	0.052	0.044
	WLS	0.0027	0.037	0.062	0.0769	0.064
75	Bootstrap[a]	0.070	0.061	0.061	0.063	0.063
	GEE2	0.046	0.058	0.057	0.051	0.040
	WLS	0.0030	0.040	0.060	0.071	0.069
100	Bootstrap[a]	0.089	0.065	0.064	0.061	0.058
	GEE2	0.057	0.054	0.050	0.053	0.040
	WLS	0.0027	0.037	0.055	0.064	0.064

[a] $q = 2000$

(DVT) detection using a multidetector-row computed tomography (MDCT) and ultrasound (US). The study also looked at the benefit of using spiral (more images and possibility of multiplanar reconstructions) with respect to sequential technique (less slices, less irradiation). Images were acquired in the spiral model (ankle to inferior vena cava) and reconstructed in 5 mm thickness slices every 5 mm, 20 mm and 50 mm. Two radiologists (one junior and one senior) assessed for each patient and each experimental setting (5/5, 5/20 and 5/50 slices) the presence of DVT. The aim of the study was to compare agreement of the different MDCT slices with the US method. Only data of the senior radiologist will be presented here (see Table 2).

Table 2. Cross-classification of DVT detection (0=absence, 1=presence) using different MDCT slices (5/5, 5/20 and 5/50 mm) and US in 107 patients by a senior radiologist (unpublished data)

	MDCT slices						
	5/5 mm		5/20 mm		5/50 mm		
US	0	1	0	1	0	1	Total
0	96	1	95	2	96	1	97
1	0	10	1	9	2	8	10
Total	96	11	96	11	98	9	107
	$\kappa_{5/5} = 0.95$		$\kappa_{5/20} = 0.84$		$\kappa_{5/50} = 0.83$		

The observed kappa coefficients ($\pm$ SE) were 0.95 ± 0.053, 0.84 ± 0.089 and 0.83 ± 0.098 for 5/5, 5/20 and 5/50 mm slices, respectively. The bootstrap approach with 2000 iterations led to a Hotelling's T^2 value of 1.46 (p=0.48) indicating no evidence of a difference between the κ coefficients at the 5% significance level. The bootstrap estimates of bias were 0.003, 0.008 and 0.009 for the 5/5, 5/20 and 5/50 mm slices, respectively. According to the rule described in Efron [9], the bias can be ignored. The differences between the κ generated by the 2000 iterations of the bootstrap are represented in Figure 1 with the 95% confidence ellipse for the difference vector $(\kappa_{5/5} - \kappa_{5/20},\ \kappa_{5/5} - \kappa_{5/50})$.

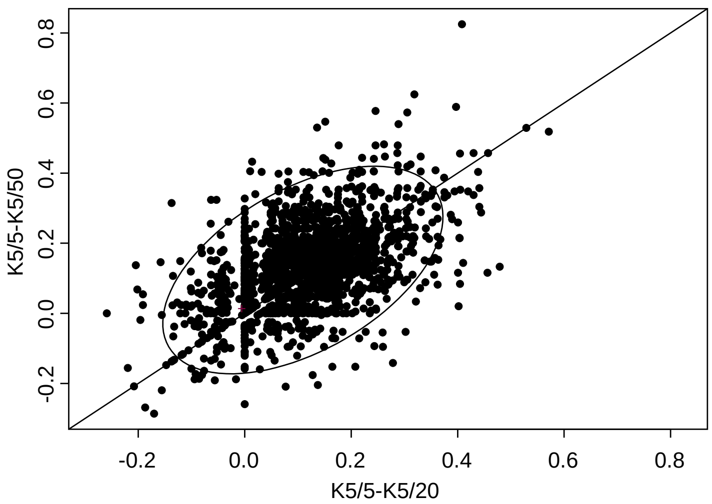

Figure 1. Kappa differences ($\kappa_{5/5} - \kappa_{5/50}$ versus $\kappa_{5/5} - \kappa_{5/20}$) generated by the bootstrap (q=2000) with 95% confidence interval.

It is seen that the origin $(0, 0)$ is well inside the confidence region, as expected.

5.2 ***Diagnosis of depression***

McKenzie et al. [3] compared for illustrative purposes the agreement between two different screening tests (Beck Depression Inventory (BDI) and General Health Questionnaire (GHQ)) and the diagnosis of depression including DSM-III-R Major depression, dysthymia, adjustment disorder with depressed mood and depression not otherwise specified (NOS). The study consisted in determining presence or absence of depression in 50 patients. Data are summarized in Table 3. McKenzie et al. found that the 95% bootstrap confidence

Table 3. Depression (0=absence, 1=presence) assessed in 50 patients according to two screening tests (BDI and GHQ) and to a medical diagnosis

	BDI		GHQ		
Depression diagnosis	0	1	0	1	Total
0	35	2	34	3	37
1	6	7	2	11	13
Total	41	9	36	14	50
	$\kappa_{BDI} = 0.54$		$\kappa_{GHQ} = 0.75$		

interval based on the percentiles for the difference between the two kappas did include 0. The kappa coefficients were 0.54 ± 0.14 between diagnosis of depression and BDI and 0.75 ± 0.11 between diagnosis of depression and GHQ, respectively. The bootstrap method described in Section 3 resulted in a T^2 value of 2.19 (p=0.14) confirming the findings of McKenzie [3]. The bootstrap estimates of bias were 0.008 and 0.009 for BDI and GHQ methods, respectively, and could be ignored. Figure 2 displays the kappa values for BDI and GHQ generated by the bootstrap method ($q = 1000$) with the corresponding 95% confidence interval.

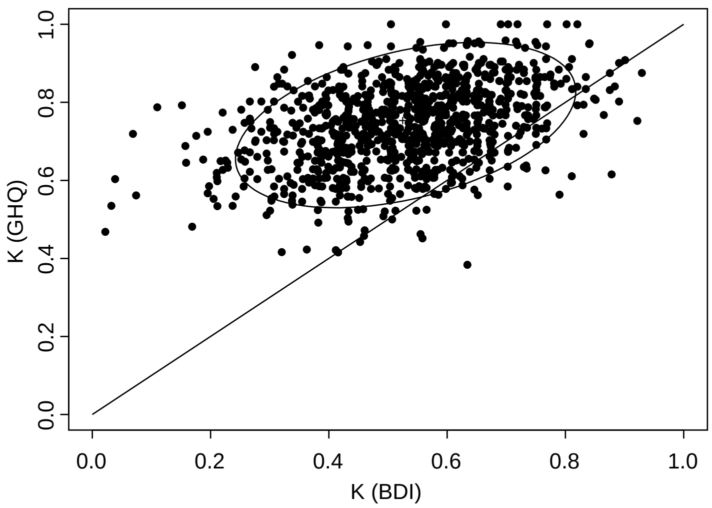

Figure 2. Kappa values of BDI and GHQ for the diagnosis of depression generated by the bootstrap ($q = 1000$) with 95% confidence interval

5.3 *Application of WLS and GEE2 approaches*

The weighted least squares method developed by Barnhart and Williamson [6] and the GEE2 approach of Williamson et al. [4] were also applied to both datasets. As seen in Table 4, these approaches led to the same conclusions as the bootstrap procedure for both examples.

6 Discussion

The comparison of two or more correlated kappa coefficients is a frequently encountered problem in real life practice and there is no simple handy test to solve it. The bootstrap method described in this work provides an estimate of

Table 4. Comparison of the bootstrap, the GEE2 and the weighted least squares (WLS) approaches applied to the radiology data (unpublished) and the depression data of McKenzie [3]

	Bootstrap			GEE2			WLS		
	κ	SE	p-value	κ	SE	p-value	κ	SE	p-value
DVT radiology data									
5/5 mm	0.95	0.056	0.48	0.95	0.048	0.56	0.95	0.053	0.46
5/20 mm	0.84	0.096		0.84	0.060		0.84	0.089	
5/50 mm	0.83	0.108		0.83	0.063		0.83	0.098	
Depression data									
BDI	0.54	0.144	0.14	0.54	0.115	0.13	0.54	0.141	0.13
GHQ	0.75	0.114		0.75	0.128		0.75	0.107	

the mean and the variance-covariance matrix of correlated kappa coefficients and hence a way to test their homogeneity by means of the Hotelling's T^2. This extension of the resampling method proposed by McKenzie et al. [3] provides an alternative to the existing advanced techniques of modeling κ coefficients. Furthermore, it can be used for the comparison of other correlated agreement or association indexes, like the intraclass kappa coefficient [7] and the weighted kappa coefficient [8] for example. The weighted least squares method developed by Barnhart and Williamson [6] and the GEE2 approach of Williamson et al. [4] led to the same conclusions as the bootstrap procedure for both examples, although estimates of the κ coefficients obtained with the bootstrap method were slightly biased. However, Efron [9] suggested that if the estimate of the bias ($\hat{bias}$) is small compared to the estimate of the standard error ($\hat{SE}$), i.e. $\hat{bias}/\hat{SE} \leq 0.25$, the bias can be ignored. Otherwise, it may be an indication that $\hat{\kappa}$ is not an appropriate estimate of the parameter κ. The bootstrap approach also yields slightly higher standard errors than the WLS and the GEE2 methods, as it was expected from the results of the simulations. Indeed, the type I errors obtained with the bootstrap method were more liberal than those with the GEE2 method, in particular if the sample size (n) was small with respect to the number (G) of kappas to be compared. This finding confirms the remark made by McKenzie [3] et al. Nevertheless, the type I error obtained by the bootstrap remains acceptable although it is recommended to use more than 1000 bootstrap iterations when the number of κ coefficients to be compared is greater than 2. The method outlined in Section 3 can be easily implemented in many statistical packages and programming languages since the method merely requires the generation of random uniform numbers and simple matrix calculations. By contrast, modeling techniques require specific programming for each problem encountered in practice. Their use is nevertheless highly recommended when it comes to account for many covariates. A function for the bootstrap method was developed in R language and is available on request from the first author.

The authors are grateful to Dr B. Ghaye, senior radiologist at the university hospital, for providing the medical imaging data.

References

[1] Cohen J., 1960, A coefficient of agreement for nominal scales., *Educational and Psychological Measurement,* **20**, 37–46.

[2] Fleiss J.L. , 1981, *Statistical methods for rates and proportions*, (2nd edn) (Wiley, New York).

[3] McKenzie D.P. et al., 1996, Comparing correlated kappas by resampling: is one level of agreement significantly different from another? *Journal of psychiatric research,* **30**, 483–492.

[4] Williamson J.M. et al., 2000, Modeling kappa for measuring dependent categorical agreement data, *Biostatistics,* **1**, 191–202.

[5] Lipsitz S.R. et al., 2001, A simple method for estimating a regression model for κ between a pair of raters, *Journal of the Royal Statistical Society A,* **164**, 449–465.

[6] Barnhart H.X. and Williamson J.M., 2002, Weighted least-squares approach for comparing correlated kappa, *Biometrics*, **58**, 1012–1019

[7] Kraemer H.C., 1979, Ramification of a population model for κ as a coefficient of reliability, *Psychometrika*, **44**, 461–472

[8] Cohen J., 1968, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological bulletin*, **70**, 213–220

[9] Efron B. and Tibshirani R.J., 1993, *An introduction to the bootstrap,* (Chapman and Hall, New York).