Int. J. Adv. Eng. Pure Sci. 2020, 32(3): 287-294
DOI: 10.7240/jeps.647755

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Inter-Rater Agreement under Stratified Random Sampling Scheme

Tabakalı Örnekleme Şeması Altında Değerlendiriciler Arası Uyum

Ayfer Ezgi YILMAZ[1], **Serpil AKTAŞ ALTUNAY**[1]

[1] *Hacettepe University, Department of Statistics, Beytepe, Ankara- Turkey.*

Abstract
Kappa coefficient is a popular statistic to measure the agreement between the classifications of two raters. The estimation of the coefficient and its variance are approximately satisfied by a simple random sampling method. To get more efficient results for confidence interval estimation of kappa, stratified random sampling method can also be used, alternatively. In this study, a bootstrap method under stratified sampling is suggested to use to estimate the confidence interval of the kappa coefficient. The results are discussed over three data sets.
Keywords: agreement, bootstrap, kappa coefficient, stratified random sampling.

Öz
Kappa katsayısı iki değerlendiricinin sınıflandırmaları arasındaki uyumu ölçen popüler bir istatistiktir. Kappa katsayısı ve varyansının tahmini yaklaşık olarak basit rasgele örnekleme yöntemiyle elde edilir. Kappa istatistiğinin güven aralığı tahmininde daha etkin sonuçlar elde edebilmek için tabakalı rasgele örnekleme yöntemi de kullanılabilir. Bu çalışmada, kappa katsayısının güven aralığı tahmin etmek için tabakalı rasgele örneklem seçimi altında bir bootstrap yöntemi önerilmiştir. Sonuçlar üç veri kümesi üzerinden tartışılmıştır.
Anahtar Kelimeler: uyum, bootstrap, kappa katsayısı, tabakalı rasgele örnekleme.

I. INTRODUCTION

The cross-classification tables where the row and column variables have the same categories are called square contingency tables. These tables are also called $R \times R$ tables and the variables of these tables are dependent. In medical or behavioral sciences, the tables occur when objects rated independently by two raters or twice by the same rater.

For these kinds of tables, the agreement between the classifications of raters (or time points) is investigated. Agreement coefficients are used to measure the level of agreement. Cohen [1] kappa coefficient as a chance-corrected measure of agreement is one of the most used statistics in literature.

In this study, we focused on estimating the confidence interval for the kappa coefficient with the bootstrap method under stratified random sampling. The bootstrap technique allows the researchers to conclude from data without any assumptions about the data or the statistic being calculated. Sometimes it is needed to estimate the precision of this statistic by resampling techniques such as bootstrap to understand its confidence interval. All of the discussed content is illustrated in three illustrative examples.

Cohen's kappa coefficient is cited in Section 2. Kappa coefficient under stratified random sampling is reviewed in Section 3. Section 4 presents the illustrative examples, followed by conclusions in Section 5.

II. COHEN'S KAPPA COEFFICIENT

Let N_{ij} denote the number of observations and N shows the total number of observations. The cell probabilities are $P_{ij} = N_{ij}/\mathrm{N}$. For each rater, the category marginal frequencies sum to one. Let N_i indicates the i*th* row total and M_i indicates the j*th* column total, for each rater. $P_{i.} = N_i/\mathrm{N}$ indicates the i*th* row total probability and $P_{.j} = N_j/\mathrm{N}$ indicates the j*th* column total probability in an $R \times R$ contingency table. Table 1 represents an $R \times R$ contingency table.

Corresponding Author: Ayfer Ezgi YILMAZ, Tel: +90 0312 297 79 00, e-posta: ezgiyilmaz@hacettepe.edu.tr
Submitted: 17.11.2019, **Revised:** 12.05.2020, **Accepted:** 04.06.2020

Table 1. $R \times R$ contingency table with N observations

Rater 1	Rater 2				
	1	2	…	R	Total
1	N_{11}	N_{12}	…	N_{1R}	N_1
2	N_{21}	N_{22}	…	N_{2R}	N_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
R	N_{R1}	N_{R2}	…	N_{RR}	N_R
Total	M_1	M_2	…	M_R	N

Inter-rater agreement is a very important context in many fields. To analyze the inter-rater agreement, the kappa coefficient is suggested by Cohen [1]. The formulas are based on multinomial sampling, which is approximately satisfied by simple random sampling. The kappa coefficient $\hat{\kappa}$ is

$$\hat{\kappa} = \frac{P_0 - P_e}{1 - P_e} \quad (1)$$

where P_o is the observed agreement,

$$P_o = \sum_{i=1}^{R} P_{ii}, \quad (2)$$

and P_e is the proportion agreement expected by chance,

$$P_e = \sum_{i=1}^{R} P_{i.} P_{.i}. \quad (3)$$

The kappa coefficient can be rewritten with respect to the observed frequencies,

$$\hat{\kappa} = \frac{N \sum_{i=1}^{R} N_{ii} - \sum_{i=1}^{R} N_i M_i}{N^2 - \sum_{i=1}^{R} N_i M_i}. \quad (4)$$

Kappa coefficient changes between -1 to 1. When there is a complete agreement between raters $\kappa = 1$ and there is no agreement between raters κ =0. Negative values of kappa indicate agreement less than chance.

The approximate asymptotic variance of κ is [2]

$$\hat{V}(\kappa) = \frac{A + B - C}{N(1 - P_e)^2}, \quad (5)$$

where,

$$A = \sum_{i=1}^{R} P_{ii}\left(1 - (P_{i.} + P_{.i})(1 - \hat{\kappa})\right)^2,$$
$$B = (1 - \hat{\kappa})^2 \sum_{i \neq j} P_{ij}\,(P_{i.} + P_{.i})^2, \quad (6)$$
$$C = \hat{\kappa}^2 - P_e(1 - \hat{\kappa})^2.$$

The confidence interval of κ is

$$\kappa \pm Z_{\alpha/2} \times \sqrt{\hat{V}(\kappa)} \quad (7)$$

where Z_α is the percentile from a standard normal distribution appropriate for the desired confidence probability [1].

Besides simple random sampling, estimation for κ and its variance under stratified random sampling will be focused in the next section.

III. KAPPA COEFFICIENT UNDER STRATIFIED RANDOM SAMPLING

Unlike the assumption of simple random sampling scheme, Stehman [3] suggested an estimator and variance estimator for estimating the kappa coefficient under stratified random sampling (KS). Stehman [3]'s simulation results showed that the estimators have small bias, and confidence intervals perform well, often even at relatively small sample sizes, when compare to the simple random sampling. For the calculation of KS, rater 1 in Table 1 is treated as a stratum and, rater 2 as reference. The reason for this assumption is explained by Stehman [3] as

"Stratified sampling [4] is a potentially useful design for accuracy assessment. In particular, if strata are constructed on the basis of the categories of strata, stratified sampling permits control over the number of sample observations in category. This guarantees that a minimum sample size can be selected in each stratum or category."

For a stratified sampling design, the row totals N_i are assumed as known, but the column totals, M_j are assumed as unknown, and N_{ij} are also unknown. In stratified random sampling, a simple random sample of observations selected in each stratum. n_i in Table 2 are the stratum sample sizes which is selected from N_i observations for $i = 1,2,\dots,R$.

Table 2. $R \times R$ contingency table under stratified sampling

Rater 1	Rater 2				
	1	2	…	R	Total
1	n_{11}	n_{12}	…	n_{1R}	n_1
2	n_{21}	n_{22}	…	n_{2R}	n_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
R	n_{R1}	n_{R2}	…	n_{RR}	n_R
					n

The row totals n_i are fixed by the stratified design, but n_{ij} depends on the observed sample. In general n sample is selected from N observations. Stehman [3]'s estimator of κ for stratified sampling is

$$\widehat{KS} = \frac{N\sum_{i=1}^{R}\widehat{N}_{ii} - \sum_{i=1}^{R}N_i\widehat{M}_i}{N^2 - \sum_{i=1}^{R}N_i\widehat{M}_i}, \qquad (8)$$

where $\widehat{N}_{ii} = \frac{N_i}{n_i}n_{ii}$ is an unbiased estimator of N_i. For any column $j = 1,2,\ldots,R$, $\widehat{M}_j = \sum_{i=1}^{R}\frac{N_i}{n_i}n_{ij}$ is an unbiased estimator of M_j. KS is not an unbiased estimator of κ, but it is a consistent estimator and the bias of KS is shown to be small in the populations examined in the subsequent empirical study [3]. The variance of KS is

$$\bar{u}_i = \frac{1}{n_i}\left\{n_{ii}\left(\frac{N}{N^2-\hat{C}}\right) + \frac{N(\hat{D}-N)}{(N^2-\hat{C})^2}\sum_{\substack{j=1\\ j\neq i}}^{R} n_{ij}N_j\right\}, \qquad (9)$$

$$u_i^2 = n_{ii}\left\{\frac{N}{N^2-\hat{C}} + \frac{N_iN(\hat{D}-N)}{(N^2-\hat{C})^2}\right\}^2 + N^2\frac{(\hat{D}-N)^2}{(N^2-\hat{C})^4}\sum_{\substack{j=1\\ j\neq i}}^{R} n_{ij}N_j^2, \qquad (10)$$

$$\hat{V}_i = \frac{u_i^2 - n_i\,\bar{u}_i^2}{n_i - 1}, \qquad (11)$$

$$\hat{V}(\mathrm{KS}) = \sum_{i=1}^{R} N_i^2\,(1-f_i)\;\hat{V}_i/n_i, \qquad (12)$$

where $\hat{D} = \sum_{i=1}^{R}N_{ii}$, $\hat{C} = \sum_{i=1}^{R}N_i\widehat{M}_i$, and the sampling fraction in stratum i is $f_i = n_i/N_i$. For strata in which n_i is small relative to N_i, the finite population correction factor $(1-f_i)$ may be ignored. A confidence interval for KS is constructed using $\widehat{KS} \pm Z_{\alpha/2} \times \sqrt{\hat{\mathrm{V}}(\mathrm{KS})}$.

Stehman [3] calculated KS coefficient under equal allocation of samples to strata which is all n_i's are equal. In this study, we used the constraint in Equation (13) to get a stratified sampling.

$$\frac{N_i}{N} = \frac{n_i}{n} \qquad i = 1,2,\ldots,R \qquad (13)$$

IV. ILLUSTRATIVE EXAMPLES

Three real-life data sets are utilized to illustrate the performance of the kappa coefficient under simple and stratified random samples. 3×3, 4×4, and 5×5 contingency tables are used. The bootstrap estimations, the bootstrap estimation of the variances, and the bootstrap estimates of 95% confidence intervals are calculated based on 10,000 replications [6,7].

In the illustrative examples, $\hat{\kappa}$ indicates the estimates of kappa coefficient and $\hat{V}(\kappa)$ indicates its variance. $\hat{\kappa}_B$ and $\widehat{KS}$ indicate the bootstrap estimates of $\hat{\kappa}$ under simple and stratified random sampling, respectively. $\hat{V}(\kappa_B)$ and $\hat{\mathrm{V}}(KS)$ indicate the bootstrap estimation of their variance. $\hat{\mathrm{V}}(\hat{\kappa}_B)$ and $\hat{\mathrm{V}}(\widehat{KS}_B)$ are the variances of bootstrap estimation of $\hat{\kappa}_B$ and $\widehat{KS}$, respectively. $\widehat{LB}$ and $\widehat{UB}$ are the 95% lower and upper bounds of the estimated kappas.

4.1. Diagnosis of Diabetes

Table 3 represents a 3×3 contingency table, which is taken from Jiménez-Navarro et al. [8] which refers to the concordance oral glucose tolerance test in patients undergoing percutaneous coronary revascularization.

Table 3. Concordance oral glucose tolerance test in patients undergoing percutaneous coronary revascularization

Oral Glucose Tolerance Test at Revascularization	**Oral Glucose Tolerance Test at 1 Month**			
	A	B	C	N_i
A: Normal	17	2	3	22
B: Glucose Intolerance	22	10	4	36
C: Diabetes Mellitus	10	11	9	30
				88

Estimated kappa coefficient, its variance, lower, and upper bounds of the confidence interval of diagnosis of diabetes data are shown in Table 4. There is a "slight agreement" between test results [9].

Table 4. Estimated kappa coefficient of diagnosis of diabetes data

$\hat{\kappa}$	$\hat{V}(\kappa)$	$\widehat{LB}$	$\widehat{UB}$
0.146	0.0048	0.0108	0.2812

Tables 5 and 6 show bootstrap estimates of κ and KS and their variance, lower, and upper bounds of 95% confidence intervals, and the biases. The sample sizes are taken as 30, 50, 70, and 87. When the results are compared, the bootstrap estimation of variance for KS is smaller than the variance for κ for all considered sample sizes. The variances of bootstrap estimation of KS are found smaller than κ's. When the bootstrap results are compared to the classical estimation of κ's results, the bootstrap estimation of the variance of κ is smaller than the variance of classical estimation of κ for $n = 87$. However, this value of sample size is $n = 50$ and the variances gradually decrease with the increasing sample sizes for KS.

Table 5. The bootstrap estimation of $\hat{\kappa}$, $\widehat{V}(\kappa)$, 95% CI, and the biases under simple random sampling for diagnosis of diabetes data

n	$\hat{\kappa}_B$	$\widehat{V}(\hat{\kappa}_B)$	$\widehat{V}(\kappa_B)$	$\widehat{LB}$	$\widehat{UB}$	Bias $\hat{\kappa}_B$	Bias $\widehat{V}(\kappa_B)$
30	0.1445	0.0093	0.0132	-0.0785	0.3675	-0.0015	0.0084
50	0.1446	0.0037	0.0074	-0.0238	0.3131	-0.0014	0.0027
70	0.1462	0.0013	0.0052	0.0054	0.2870	0.0002	0.0004
87	0.1459	0.0001	0.0041	0.0203	0.2716	-0.0001	-0.0007

Table 6. The bootstrap estimation of $\widehat{KS}$, $\widehat{V}(KS)$, 95% CI, and the biases under stratified random sampling for diagnosis of diabetes data

n	$\widehat{KS}$	$\widehat{V}(\widehat{KS}_B)$	$\widehat{V}(KS)$	$\widehat{LB}$	$\widehat{UB}$	Bias $\widehat{KS}$	Bias $\widehat{V}(KS)$
30	0.1463	0.0090	0.0066	-0.0120	0.3045	0.0003	0.0018
50	0.1456	0.0035	0.0025	0.0468	0.2443	-0.0004	-0.0022
70	0.1461	0.0011	0.0008	0.0910	0.2012	0.0001	-0.0040
87	0.1459	0.0000	0.0000	0.1334	0.1583	-0.0001	-0.0047

The widths of the 95% confidence intervals of κ and KS are calculated and summarized in Figure 1. For all sample sizes, the widths of confidence interval (CI) for KS are narrower than the widths of CI for κ. The deviation of widths for KS is also smaller. While the sample size increases, the difference between the widths of these two coefficients increases.

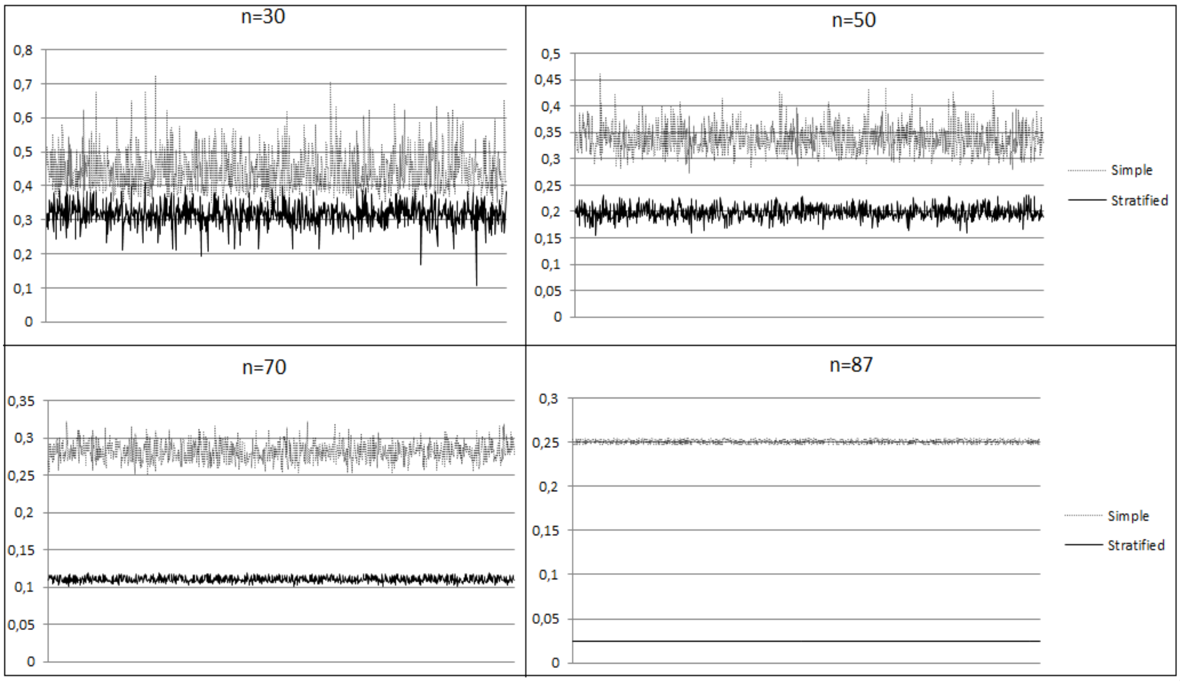

Figure 1. The widths of the 95% CI for κ and KS for diagnosis of diabetes data

4.2. Left Eye-Right Eye

The unaided distance vision data in Table 7 is taken from Stuart [10]. Data on unaided distance vision of 3242 men employed in Royal Ordnance factories in Britain from 1943 to 1946 are used. From the estimated κ in Table 8, there is a "moderate agreement" between left-right eye grades [9].

Table 7. Unaided distance vision data of 3242 men in Britain

Right Eye Grade	Left Eye Grade				
	Best	Second	Third	Worst	N_i
Best	821	112	85	35	1053
Second	116	494	145	27	782
Third	72	151	583	87	893
Worst	43	34	106	331	514
					3242

Table 8. Estimated kappa coefficient of left-right eye data

$\hat{\kappa}$	$\hat{V}(\kappa)$	$\widehat{LB}$	$\widehat{UB}$
0.574	0.0001	0.5524	0.5956

The bootstrapping results are shown in Tables 9 and 10. The sample sizes are taken as: 50, 100, 200, 500, 1000, and 2000. When the results in Tables 9 and 10 are compared, the bootstrap estimation of KS is smaller than κ for all sample sizes. The bootstrap estimations of variance and the variance values of bootstrap estimations of KS are also smaller than κ's results. When the bootstrap results are compared to the classical estimation of κ's results, the bootstrap estimation of the variance of κ is greater than the variance of classical estimation of κ for all sample sizes. However, $\hat{V}(KS)$ is smaller than the classical estimation of $\hat{V}(\kappa)$ on and after $n = 1000$. The widths of the 95% CI for κ and KS are calculated and summarized in Figure 2. The results are similar to Figure 1.

Table 9. The bootstrap estimation of $\hat{\kappa}$, $\hat{V}(\kappa)$, 95% CI, and the biases under simple random sampling for left eye-right eye data

n	$\hat{\kappa}_B$	$\hat{V}(\hat{\kappa}_B)$	$\hat{V}(\kappa_B)$	$\widehat{LB}$	$\widehat{UB}$	Bias $\hat{\kappa}_B$	Bias $\hat{V}(\kappa_B)$
50	0.5711	0.0079	0.0069	0.4090	0.7332	-0.0029	0.0068
100	0.5717	0.0038	0.0034	0.4570	0.6863	-0.0023	0.0033
200	0.5738	0.0019	0.0017	0.4927	0.6550	-0.0002	0.0016
500	0.5738	0.0007	0.0007	0.5224	0.6251	-0.0002	0.0006
1000	0.5744	0.0003	0.0003	0.5380	0.6107	0.0004	0.0002
2000	0.5744	0.0001	0.0002	0.5488	0.6001	0.0004	0.00005

Table 10. The bootstrap estimation of $\widehat{KS}$, $\hat{V}(KS)$, 95% CI, and the biases under stratified random sampling for left eye-right eye data

n	$\widehat{KS}$	$\hat{V}(\widehat{KS}_B)$	$\hat{V}(KS)$	$\widehat{LB}$	$\widehat{UB}$	Bias $\widehat{KS}$	Bias $\hat{V}(KS)$
50	0.4619	0.0059	0.0038	0.3417	0.5818	-0.1121	0.0037
100	0.4635	0.0029	0.0019	0.3793	0.5477	-0.1105	0.0017
200	0.4609	0.0014	0.0009	0.4023	0.5195	-0.1131	0.0008
500	0.4629	0.0005	0.0003	0.4277	0.4980	-0.1111	0.0002
1000	0.4615	0.0002	0.0001	0.4390	0.4839	-0.1125	0.0001
2000	0.4621	0.0001	0.00003	0.4503	0.4740	-0.1119	-0.00008

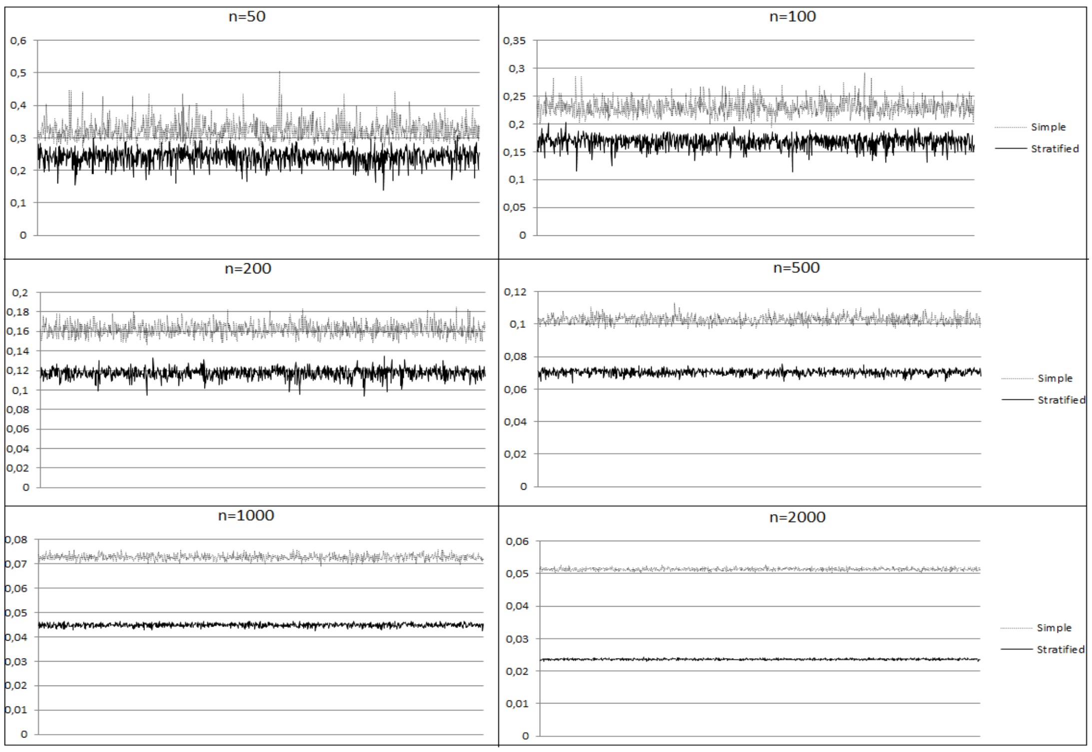

Figure 2. The widths of the 95% CI for κ and KS for left eye-right eye data

4.3. Blight Data

Table 11 shows a 5×5 table with 9660 observations which are generated randomly by Stehman [3]. From the estimated κ in Table 12, there is a substantial agreement" between strata and reference [9].

Table 11. Blight data

Strata	Reference					N_i
	1	2	3	4	5	
1	4440	0	30	30	30	4530
2	30	1500	180	0	0	1710
3	240	450	1170	180	0	2040
4	60	90	210	750	30	1140
5	0	0	30	30	180	240
						9660

Table 12. Estimated kappa coefficient of blight data

$\hat{\kappa}$	$\hat{V}(\kappa)$	$\widehat{LB}$	$\widehat{UB}$
0.754	0.00003	0.7442	0.7638

The bootstrapping results are shown in Tables 13 and 14. The sample sizes are taken as 150, 500, 1000, 2000, 3000, and 5000. The bootstrapping results are similar to previous tables. The bootstrap estimation, the variance values of bootstrap estimations, and bootstrap estimations of the variance of KS are smaller than κ for all considered sample sizes. When the bootstrap results are compared to the classical estimation of κ's results, the bootstrap estimation of the variance of κ is greater than the variance of classical estimation of κ for all sample sizes. However, $\hat{V}(KS)$ is smaller than the classical estimation of $\hat{V}(\kappa)$ on and after $n = 3000$.

The widths of the 95% CI for and are calculated and summarized in Figure 3. The results are similar to Figures 1 and 2.

Table 13. The bootstrap estimation of $\hat{\kappa}$, $\hat{V}(\kappa)$, 95% CI, and the biases under simple random sampling for blight data

n	$\hat{\kappa}_B$	$\hat{V}(\hat{\kappa}_B)$	$\hat{V}(\kappa_B)$	$\widehat{LB}$	$\widehat{UB}$	Bias $\hat{\kappa}_B$	Bias $\hat{V}(\kappa_B)$
150	0.7532	0.0018	0.0029	0.6479	0.8585	-0.0008	0.0029
500	0.7539	0.0005	0.0009	0.6964	0.8114	-0.0001	0.0008
1000	0.7548	0.0003	0.0004	0.7140	0.7955	0.0008	0.0004
2000	0.7543	0.0001	0.0002	0.7256	0.7831	0.0003	0.0002
3000	0.7544	0.0001	0.00014	0.7309	0.7779	0.0004	0.00012
5000	0.7545	0.0000	0.00008	0.7363	0.7726	0.0005	0.00006

Table 14. The bootstrap estimation of $\widehat{KS}$, $\hat{V}(\widehat{KS})$, 95% CI, and the biases under stratified random sampling for blight data

						Bias	
n	$\widehat{KS}$	$\hat{V}(\widehat{KS}_B)$	$\hat{V}(\widehat{KS})$	$\widehat{LB}$	$\widehat{UB}$	$\widehat{KS}$	$\hat{V}(\widehat{KS})$
150	0.6390	0.0010	0.0001	0.5776	0.7005	-0.1150	0.0001
500	0.6414	0.0003	0.0003	0.6081	0.6747	-0.1126	0.0003
1000	0.6402	0.0001	0.0001	0.6173	0.6630	-0.1138	0.0001
2000	0.6402	0.0001	0.00006	0.6250	0.6554	-0.1138	0.00003
3000	0.6400	0.0000	0.00003	0.6285	0.6516	-0.1140	0.00001
5000	0.6404	0.0000	0.00001	0.6329	0.6479	-0.1136	-0.00001

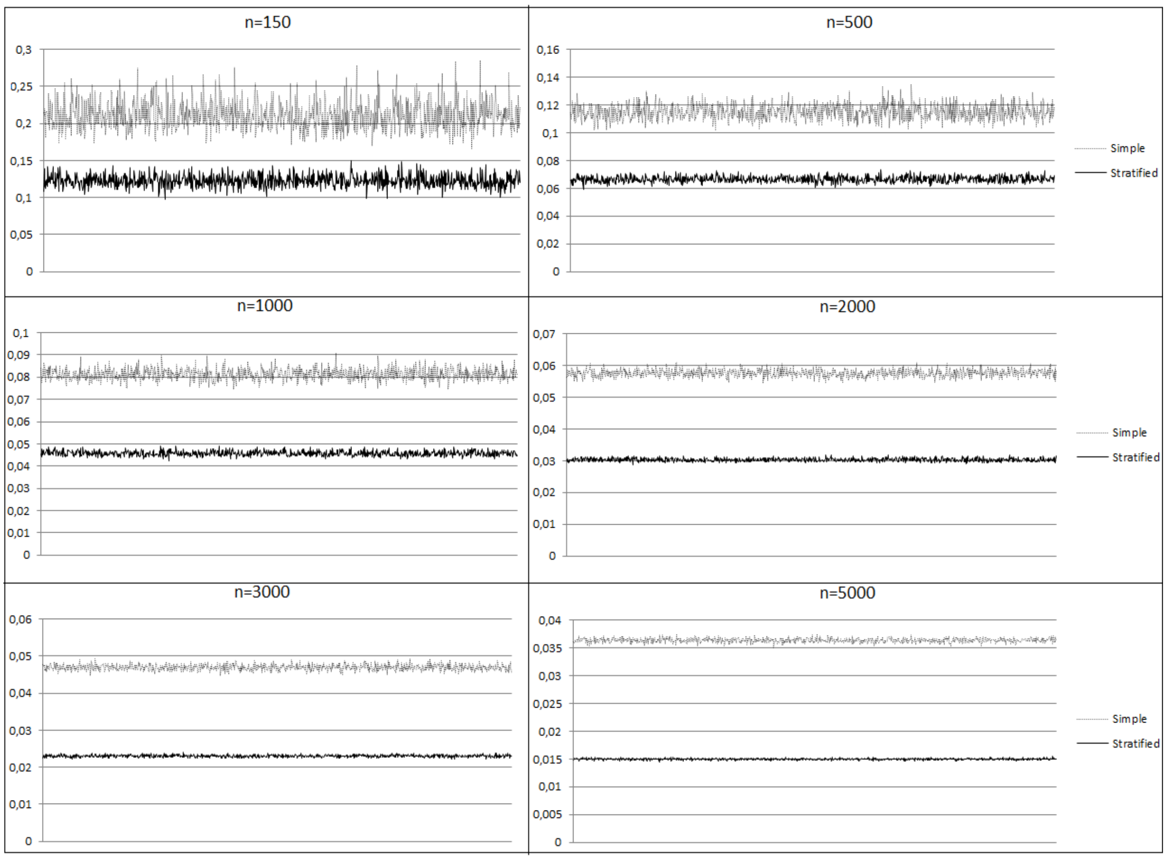

Figure 3. The widths of the 95% CI for κ and $\widehat{KS}$ for blight data

V. CONCLUSION

In recent studies, inter-rater agreement analysis has been growing extensively. There are different ideas between researchers for agreement between raters' decisions. The main argument of the researchers who prefer to use agreement models reveals pure agreement. In this paper, we present κ coefficient under simple and stratified random samplings for the study of inter-rater agreement. We focus on estimating the kappa coefficient and its variance in terms of the bootstrapping method.

One of the advantages of calculating κ under stratified random sampling is that, after the sampling procedure, the method does not allow a table in which all cells of a row are zero. However, it cannot be guaranteed for simple random sampling.

Two important characteristics of κ under stratified random sampling used as point estimates of parameters are bias and sampling variability. It is well known that bias refers to whether an estimator tends to either over or underestimates the parameter and sampling variability refers to how much the estimate varies between samples.

In the light of the two characteristics, we also calculated the biases. It is found that the biases of the estimators for entire data sets are quite small. This point to the accuracy of parameter estimation.

The estimated κ under stratified random sampling tends to give a lower agreement than simple random sampling and classical estimation of κ. The bootstrap estimation of κ under stratified random sampling has a smaller variance than simple random sampling $\left(\hat{\mathrm{V}}(\widehat{KS}_B) < \hat{\mathrm{V}}(\hat{\kappa}_B)\right)$. The bootstrap estimation of variance under stratified random sampling is smaller than simple random sampling $\left(\hat{\mathrm{V}}(\widehat{KS}) < \hat{\mathrm{V}}(\hat{\kappa}_B)\right)$. The bootstrap estimation of variance under stratified random sampling is smaller than the classical estimation of κ for some values of n.

When the sample size increases, the width of confidence interval converges to 0, and the dispersion of width decreases. For all values of n, the width and dispersion of confidence interval of κ under stratified random sampling are narrower than simple random sampling. Utilizing a stratified sampling method for confidence interval estimation of kappa statistic can lead to a more efficient statistical estimate than those of simple random samples. Calculating κ under stratified sampling provides much information over sample random sampling.

REFERENCES

[1] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37-46.

[2] Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969). Large Sample Standard Errors of Kappa and Weighted Kappa. *Psychological Bulletin*, 72(5), 323-327.

[3] Stehman, S.V. (1996). Estimating the Kappa Coefficient and Its Variance under Stratified Random Sampling. *Photogrammetric Engineering and Remote Sensing*, 62(4), 401-407.

[4] Cochran, W.G. (1977). Sampling Techniques, 3rd ed., John Wiley and Sons, New York.

[5] Särdal, C.E., Swensson, B., & Wretman, J. (1992). Model-assisted survey sampling, Springer-Verlag, New York.

[6] Efron, B. & Tibshirani R.J. (1993). An introduction to the bootstrap, Monographs on statistics and applied probability, no. 57., Chapman and Hall, London.

[7] Klar, N., Lipsitz, S.R., Parzen, M., & Leong, T. (2002). An Exact Bootstrap Confidence Interval for κ in Small Samples. *The Statistician*, 51(4), 467-478.

[8] Jiménez-Navarro, M.F., Garcia-Pinilla, J.M., & Garrido-Sanchez, L., et al. (2010). Poor Reproducibility of the Oral Glucose Tolerance Test in the Diagnosis of Diabetes during Percutaneous Intervention. *International Journal of Cardiology*, 142(3), 245-249.

[9] Landis, J.R. & Koch, G.G. (1977). The Measurement of Observed Agreement for Categorical Data. *Biometrics*, 33(1), 159-174.

[10] Stuart, A., (1953). The Estimation and Comparison of Strengths of Association in Contingency Tables. *Biometrika*, 40(1/2), 105-110.