Bilge International Journal of Science and Technology Research

© 2016 Kutbilge Akademisyenler Derneği
ISSN: 2651-401X
e-ISSN: 2651-4028
Research Article / Araştırma Makalesi

2021, Volume: 5, Issue: 2, 107-110
Received: 10.06.2021; Accepted: 08.07.2021
DOI: 10.30516/bilgesci.950797

Analysis of the Multi-raters Agreement with Log-Linear Models

Gökçen Altun [1*]

Abstract: In this study, 87 digital panoramic images are classified by the three raters to assess the accuracy of diagnosis of peri-implant bone defects. The coefficient of kappa is obtained as 0.81 among the three raters, which indicates an almost perfect agreement. Then, the log-linear agreement models are applied to the data. The best model is determined based on the model selection criteria. Using the best model, we estimate the agreement parameter. It is 33 times higher for three raters to make the same decision than to make a different decision. The results show that the coefficients of the agreement only show the value of the fit between raters. On the other hand, agreement models provide a model equation for the raters, and more detailed and consistent results can be obtained by calculating the agreement and association parameters.

Keywords: Agreement, kappa coefficient, log-linear models, peri-implant bone effect.

[1]**Address:** Bartın University, Department of Computer Technology & Information Systems,
Bartın, Turkey.

***Corresponding author**: gokcenefendioglu@gmail.com

Citation: Altun, G. (2021). Analysis of the Multi-raters Agreement with Log-Linear Models. Bilge International Journal of Science and Technology Research, 5(2): 107-110.

1. INTRODUCTION

The reliability of the measurements taken by the clinician or the treatment tool is the basis for the effective delivery of health services. Whether evaluations and findings recorded during clinical assessments are recorded by the same clinician at different times or by different clinicians over a short period, the result must be consistent. Consistency of evaluations reflects compliance. Compatibility measure; It is a measure of the consistency of two or more clinicians' evaluations about a patient or the consistency of a clinician's evaluations at different times (Gail and Benichou, 2000). It is known that there are inconsistencies and problems in many subjects in the medical sciences. In measurements; There may be differences depending on the specific sensitivities of the medical devices, the training and skills of the evaluators using the device, and the characteristics of the units concerned. Therefore, differences that may occur in diagnosis pose a problem (Broemeling, 2007). The compatibility studies between evaluations of multiple decision-makers, experts and diagnostic tests are frequently encountered in many areas. Categorical evaluations of interest; binary classification (yes / no, etc.), ordinal (low, medium, high) and nominal (schizophrenic, manic depression, severe depression, etc.) evaluations (Uebersax, 1992). In such studies, it is very important to investigate whether there is a statistical agreement between those who evaluate a situation. The number of raters can be more than two. They are called multi-raters. (Saraçbaşı, 2011). The kappa coefficient of Cohen (1960) is used to measure the agreement between 2 raters, as follows

$$\kappa = \frac{\sum_{i=1}^{R} p_{ii} - \sum_{i=1}^{R}\sum_{j=1}^{R} p_{i.}p_{.j}}{1 - \sum_{i=1}^{R}\sum_{j=1}^{R} p_{i.}p_{.j}} \qquad (1)$$

For the contingency tables, p_{ij} represents the probability that an observation fall in the ith row and jth column, $p_{i.}$ and $p_{.j}$ denote the marginal probability of the table. If the row and column classifications are ordinal, the weighted kappa is preferred. The weighted kappa is calculated by

$$\kappa = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{i.}p_{.j}}{1 - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{i.}p_{.j}} \qquad (2)$$

where w_{ij} is weight range $0 \leq w_{ij} \leq 1$ (Agresti, 2002). The weight of Fleiss-Cohen-Everitt (1969) is $w_{ij} = 1 - |i-j|/R$ and the weight of Fleiss-Cohen (1973) is $w_{ij} = 1 - (i-j)^2/(R-1)^2$.

Kendall's agreement coefficient is used to assess compatibility between more than two raters on the ordinal scale. Kendall W takes values from 0 to 1. It is a measure of compatibility between p raters that evaluate n people. There are two ways to calculate Kendall W. First, the row marginal sums of ranks are obtained according to individuals and calculated from R_i to S or S' (Kendall et.al., 1939, Landis et.al., 1977, Lawal, 2003, Saraçbaşı, 2011, Siegel, 1956), given by

$$S = \sum_{i=1}^{n}(R_i - \bar{R})^2 \text{ or } S' = \sum_{i=1}^{n} R_i^2 = SSR \qquad (3)$$

where S is the sum of squares over the row totals of ranks, $\bar{R}$ is the mean of the R_i values. The Kendall W can be derived by using the below equations

$$W = \frac{12S}{p^2(n^3-n)-pT} \text{ or } W = \frac{12S' 3p^2n(n+1)^2}{p^2(n^3-n)-pT} \qquad (4)$$

where t_k is the number of equivalent ranks within each of the m groups containing equivalent evaluations. The correction factor T for equivalent ranks is calculated as follows

$$T = \sum_{k=1}^{m}(t_k^3 - t_k) \qquad (5)$$

These results were interpreted according to the criteria of Landis and Koch as κ < 0 poor agreement; κ = 0–0.20 slight agreement; κ = 0.21–0.40 fair agreement; κ = 0.41–0.60 moderate agreement; κ = 0.61–0.80 substantial agreement; and κ = 0.81–1.00 almost perfect agreement (Landis and Koch, 1977). Although the Kappa coefficient is a widely and popularly used coefficient, studies have been conducted to examine its advantages and disadvantages. (Tanner and Young MA, 1985a and 1985b). The kappa statistic reduces all information about the agreement to a single number. On the contrary, the agreement models give more detailed information about the study and have many advantages (Broemeling, 2007). In the study, the evaluation of fit models was made. Kappa coefficient was calculated for 87 panoramic images handled by three raters. The results for the calculated coefficient and model evaluation are discussed.

2. MATERIAL VE METHOD

2.1 Log-linear Models

Log-linear models for contingency tables are similar in concept to the analysis of variance used for the factor-response variable. The difference between them is that in the analysis of variance, the response variable is normally distributed continuous variables, while in the log-linear models, the response variable is assumed to be Poisson distributed. (Uebersax,1992).

In the case of more than two categorical variables, the use of chi-square independence tests in the determination of the relationship between the variables in the contingency tables becomes difficult or sometimes impossible. In this case, logarithmic linear models, which allow the testing of a much larger number of hypotheses compared to the chi-square, which do not impose restrictions on the number of rows and columns in both the two-dimensional tables where the chi-square can be applied, and the three-dimensional tables where the chi-square is insufficient, is preferred. In the multi-dimensional contingency tables in logarithmic linear models, a model is created to investigate the relationships between the variables. The parameters in the model are estimated and the significance of this model is tested. The goodness of fit of a model is the evaluation of observed and expected frequencies by comparing them. Likelihood ratio statistics (G^2) and Pearson chi-square test statistics are frequently used goodness-of-fit test statistics. (Agresti, 2002).

2.2 Agreement Models

Now, we defined the log-linear models for three-dimensional contingency tables (Uebersax, 1992).

Model:

1.$\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + I(i = j) + I(j = k) + I(j = k) + I(i = j = k)$ (6)

2. $\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \beta^{AB}u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + I(i = j)\ I(j = k) + I(j = k) + I(i = j = k)$ (7)

3. $\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \beta^{AB}u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + \beta^{ABC}u_iv_jw_k$ (8)

4. $\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \beta^{AB}u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + I(i = j) + I(j = k) + I(j = k)$ (9)

5. $\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \beta^{AB}u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + I(i = j = k)$ (10)

6. $\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \beta^{AB}u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + \beta^{ABC}u_iv_jw_k + I(i = j) + I(j = k) + I(j = k)$ (11)

7. $\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \beta^{AB}u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + \beta^{ABC}u_iv_jw_k + I(i = j) + I(j = k) + I(j = k) + I(i = j = k)$ (12)

If the terms of the models will be explained; m_{ijk} is the expected frequency. The parameter λ shows the constant term. The parameter λ_i^A shows the effect of ith decision of rater A. The parameter λ_i^B shows the effect of ith decision of rater B. The parameter λ_i^C shows the effect of ith decision of rater C where i,j,k=1,…R and R represents the rating category. The below constraints should be hold

$$\sum_{i=1}^{R}\lambda_i^A = \sum_{j=1}^{R}\lambda_j^B = \sum_{k=1}^{R}\lambda_k^C = 0,$$

where β^{AB} ,β^{AC}, β^{BC} are association parameters between two evaluators. However, the parameter β^{ABC} is the association parameter between three evaluators. u_i, v_j and w_k are respectively the score values that belong to evaluators A, B, and C. They are defined as $u_i = i$ for rater A; $v_j = j$ for rater B; $w_k = k$ for rater C. The parameters $I(i = j), I(j = k)$ and $I(j = k)$ are agreement parameters between two evaluators. However, $I(i = j = k)$ is the agreement parameter between three r evaluators (Saraçbaşı, 2011).

2.3 Data Analysis

2.3.1 Data

The models introduced in the previous section are applied on a real data set. The research protocol was approved by the Ethical Committee for Animal Research of the Ordu University with the assignment protocol 2016/14. The images were evaluated separately by three raters, each of whom had at least 10 years of experience in implant surgery or imaging applications. The raters scored the images on a five-point Likert scale asking whether a peri-implant bone defect was (1) definitely absent, (2) probably absent, (3) unsure, (4) probably present, (5) definitely present. There are too many sample zeros in the original study. For this reason, the categories (1)-(2) and (4)-(5) have been combined and the number of levels has been reduced. The results belonging to the raters are given in Table 1.

Table 1. The results of 87 digital panoramic images according to 3 raters

A	B	C		
		1	2	3
1	1	33	4	1
	2	0	0	3
	3	0	2	1
2	1	0	1	0
	2	0	3	0
	3	0	0	0
3	1	3	0	1
	2	0	2	0
	3	3	2	28

3. RESULTS

The weighted kappa coefficient is calculated as 0.70 between rater A and B, 0.74 between rater A and C, and 0.58 between rater B and C. According these results, we conclude that the agreement between the A-B and A-C raters are substantial, while the agreement between the B-C rater is moderate. The Kendall's coefficient of congruence, calculated for the fit between three raters, is obtained as 0.81. So, the agreement between the three raters is substantial. The models given in Section 2.2 are fitted to the current data and the results are given in Table 2.

Table 2. The goodness of fit statistics of models and information criteria

Models	G^2	df	p-value	AIC
1	22.246	16	0.135	-9.754
2	16.919	13	0.203	-9.081
3	43.251	16	<0.000	-
4	21.959	14	0.079	-6.041
5	18.691	16	0.285	-13.309
6	21.641	13	0.061	-4.359
7	16.595	12	0.165	-7.405

All models, except 3 provides accurate fits to the data set. To select the best fitted model, the Akaike Information Criterion $(AIC = G^2 - 2df)$ is calculated. It is the best model with the smallest AIC value. According to this rule, model 5 is selected as a best model. The parameter estimates and odds ratios are calculated for model 5 and given in Table 3.

Table 3. Parameter estimates and odds ratio values of Model 5

Parameter	Estimation	St. Error	Z-value	Odds Ratio
β^{AB}	0.386	0.299	1.292	1.471
β^{AC}	-0.775	0.403	-1.925**	2.171
β^{BC}	0.324	0.388	0.836	1.383
$I(i = j = k)$	3.502	0.842	4.158*	33.182

$^{*}p<0.05$ $^{**}p<0.10$

According to the results in Table 3; The highest relationship between raters is between A and C, while the lowest is between rater B and rater C. Based on the odds ratios from Table 3, the probability of giving i+1 decision rather than i of rater C is 2 times higher than giving i+1 decision rather than i of rater A. 3 raters are 33 times more likely to make the same decision than they are to make a different decision.

4. DISCUSSION

For nominal variables, agreement between raters is measured by the kappa coefficient. If the variables are ordinal, it is more appropriate to use the weighted kappa coefficient. The agreement between 3 or more raters is measured by Kendall's agreement coefficient. However, since these calculated coefficients are a single number, they are not sufficient for a detailed interpretation. For this reason, in addition to calculating kappa coefficient, agreement analysis with log-linear models has become widespread. In studies where the scale is in order, it is important to use log-linear models to evaluate agreement and association separately. In these models, the association and agreement parameters can be estimated and interpreted separately. Odds ratios are obtained with the calculated parameters and allow the interpretation of the relationship. In this study, seven different models that examine agreement and association separately and together are introduced. Models were implemented for an agreement between more than two evaluators. In other words, this study offers that researchers can make a more detailed interpretation about modeling the agreement between raters and their work with the obtained parameters.

REFERENCES

Agresti A. Categorical data analysis. New York: Wiley; 2002.

Broemeling, Lyle D. Bayesian biostatistics and diagnostic medicine. CRC press, 2007.

Cohen JA. Coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960; 20: 37-46.

Fleiss J, Cohen J, Everitt, BS. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin 1969; 72: 323-7.

Fleiss J, Cohen J. The equivalence of weighted kappa and intraclass correlation coefficient as measure of reliability. Educational and psychological measurement 1973; 33: 613-9.

Gail M H, Benichou J. Encylopedia of Epidemiologic Methods. 1 st .Ed., New York: Wiley, 2000: 35-47.

Kendall M G, Babington-Smith B. The Problem of m Rankings. The Annals of Mathematical Statistics, 1939; 10 (3): 275- 287.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74.

Lawal, HB., Categorical Data Analysis with SAS and SPSS Application. Lawrence ERlbawn Association Publisher, London, 2003

Saraçbaşi, Tülay. "Agreement models for multiraters." Turkish Journal of Medical Sciences 41.5 (2011): 939-944.

Siegel S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw- Hill, 1956.

Tanner MA, Young MA. Modeling agreement among raters. Journal of American Statistical Association 1985a; 80: 175-80.

Tanner MA, Young MA. Modeling ordinal scale disagreement. Psychological Bulletin 1985b; 98: 408-15.

Uebersax, J. S. "Modeling approaches for the analysis of observer agreement." Invest Radiol 27.9 (1992): 738-743.