Original Article

Turk J Med Sci
2011; 41 (5): 939-944
© TÜBİTAK
E-mail: medsci@tubitak.gov.tr
doi:10.3906/sag-1006-891

Agreement models for multiraters

Tülay SARAÇBAŞI

Aim: Agreement between 2 or more independent raters evaluating the same items and same scale can be measured by kappa coefficient. In recent years, modeling agreement among raters rather than summarizing indices has been preferred. In this study, the disadvantages of kappa are reviewed. Agreement models are introduced and these models are applied to a real data set.

Materials and methods: Three pathologists classified each of 118 slides in terms of carcinoma in situ of the uterine cervix, based on the most involved lesions. Using log-linear agreement models, agreement between 3 pathologists according to their evaluations was investigated.

Results: Coefficient of kappa was found to be 0.48 among the 3 pathologists, which indicates a moderate agreement. Models were applied to the data. The agreement parameter was estimated for the best model among models. The probability of giving the same decision by the 3 pathologists was 2.5 times higher than that of giving a different decision.

Conclusion: Log-linear models can be used to measure the agreement among more than 2 raters. Modeling agreement can provide more information than kappa.

Key words: Agreement, log-linear models, uterine cancer

Çoklu değerlendiriciler için uyum modelleri

Amaç: Aynı birimleri aynı ölçeğe göre değerlendiren iki ya da daha fazla değerlendirici arasındaki uyum kappa katsayısı ile ölçülebilir. Son yıllarda kappa katsayısı ile uyumu özetlemek yerine değerlendiriciler arasındaki uyumun modellenmesi tercih edilmektedir. Bu çalışmada, kappa katsayısının olumsuz yönlerine değinilmiş ve uyum modelleri tanıtılmıştır. Ayrıca bu modeller gerçek bir veri kümesi üzerinde uygulanmıştır.

Yöntem ve gereç: Yüzonsekiz slayt, rahimde kanser olup olmadığını incelemek için içerdikleri lezyonlara göre üç patolog tarafından sınıflandırılmıştır. Kurulan log-doğrusal modeller ile üç patolog arasında verdikleri karar açısından uyum olup olmadığı araştırılmıştır.

Bulgular: Üç patolog arasındaki uyum, kappa uyum katsayısına göre 0,48, yani orta uyum bulunmuştur. Veriler uyum modelleriyle incelenmiş ve en iyi uyumu gösteren model için uyum parametresi tahmin edilmiştir. Üç patologun aynı kararı verme olasılığı, farklı karar verme olasılığının yaklaşık 2,5 katı bulunmuştur.

Sonuç: Log-doğrusal modeller yardımıyla ikiden fazla değerlendirici arasındaki uyum incelenebilir. Bu modeller ile daha ayrıntılı ve tutarlı sonuçlara ulaşılır.

Anahtar sözcükler: Uyum, log-doğrusal modeller, rahim kanseri

Received: 14.06.2010 – Accepted: 22.09.2010

Department of Statistics, Faculty of Science, Hacettepe University, Beytepe, Ankara - TURKEY

Correspondence: *Tülay SARAÇBAŞI, Department of Statistics, Faculty of Science, Hacettepe University, Beytepe, Ankara - TURKEY*
E-mail: toker@hacettepe.edu.tr

Introduction

In medicine, behavioral sciences, and education, the agreement between different raters for the same subject has been widely investigated. The clinical agreement in some medical fields, such as pathology and psychology must be consistent. For example, with the help of magnetic resonance and ultrasonic visualization techniques, the stage of cancer can be detected. There should be consistency between cellular diagnosis and pathological diagnosis used in diagnosing cancer. When the samples taken from the patients are evaluated in different laboratories in order to understand if there is an illness, an agreement between laboratories is expected. In these kinds of studies, the search of whether there is a statistical agreement between different raters evaluating the same fact gains importance. If the number of the rater is more than 2, it is called multirater. The agreement between 2 raters is measured by kappa coefficient developed by Cohen (1) and estimated as

$$\kappa = \frac{\sum_{i=l}^{R} \mathrm{P}_{ii} - \sum_{i=i}^{R}\sum_{j=l}^{R} \mathrm{P}_{i}.\mathrm{P}._{j}}{1 - \sum_{i=i}^{R}\sum_{j=l}^{R} \mathrm{P}_{i}.\mathrm{P}._{j}} \qquad (1)$$

In the R × R square contingency tables for 2 raters, p_{ij} denotes the probability of assigning an item response i for the first rating and second rating, $p_{i.}$ and $p._{j}$ as the marginal probability of assigning an item response i for the first rating and j for the second rating where R is the number of rating nominal scale. When the row and column classifications are ordinal, then weighted kappa is used,

$$\kappa = \frac{\sum_{i=l}^{R}\sum_{j=l}^{R} \mathrm{w}_{ij}\mathrm{P}_{ij} - \sum_{i=l}^{R}\sum_{j=l}^{R} \mathrm{w}_{ij}\mathrm{P}_{i}.\mathrm{P}._{j}}{1 - \sum_{i=l}^{R}\sum_{j=l}^{R} \mathrm{w}_{ij}\mathrm{P}_{i}.\mathrm{P}._{j}} \qquad (2)$$

where w_{ij} is the weight range $0 \leq \mathrm{w}_{ij} \leq 1$ (2). Fleiss -Cohen-Everitt weight (3) is $\mathrm{w}_{ij} = 1 - |i - j|/R$ and Fleiss-Cohen weight (4) is $\mathrm{w}_{ij} = 1 - (i-j)^2/(R-1)^2$. Landis and Koch (5) defined the agreement levels of kappa coefficient as: <0 poor, 0-0.2 slight, 0.2-0.4 fair, 0.4-0.6 moderate, 0.6-0.8 substantial, and 0.8-1 almost perfect.

Although it is really popular and widely used, the advantages and disadvantages of kappa coefficient have been argued. Based on these discussions, new approaches have been presented regarding agreement (6,7). For instance, kappa is mostly dependent on the actual prevalence of the diagnosed event. While the sensitivity and specificity rates of some tests used in diagnosing some markers are high, their accuracy can be lower as a result of the low level of the prevalence of the illness. Consequently, dependence of kappa on prevalence creates difficulty. Another disadvantage of kappa occurs in unbalanced marginal totals. Kappa gained in this case is larger than that gained from balanced marginal totals. In this study agreement models for multiraters are dealt with, these models and kappa coefficient for 118 patient evaluated by 3 raters is calculated, and the results are discussed.

Materials and methods

Log-linear models

The log-linear model is one of the specialized cases of generalized linear models for Poisson and multinomial data. Log-linear analysis is an extension of the R × C contingency table where the conditional relationship between 2 or more discrete, categorical variables is analyzed by taking the natural logarithm of the expected frequencies for the given model obtained within a contingency table. Although log-linear models can be used to analyze the interaction between 2 nominal categorical variables (2-way contingency tables), they are more commonly used to evaluate multiway contingency tables that involve 3 or more variables (2).

Linear-by-linear association models can be used to analyze the association between 2 or more ordinal categorical data. Row effect models can be used to analyze the association between nominal row variables and ordinal column variables (2).

The overall goodness-of-fit of a model is assessed by comparing the expected frequencies to the observed cell frequencies for each model. The Pearson chi-square statistic or the likelihood ratio statistic (G^2) can be used to test a model fit. G^2 is more commonly used because it is the statistic that is minimized in maximum likelihood estimation (2)

Agreement models

In this section log-linear models for a 3 dimensional contingency table showing the results of 3 raters are defined (8). In these models R is the rating categories.

Model:

$$1.\log(m_{ijk}) = \lambda + \lambda^A_i + \lambda^B_j + \lambda^C_k + I\,(i = j) + I\,(i = k) + I\,(j = k) + I\,(i = j = k) \quad (3)$$

$$2.\log(m_{ijk}) = \lambda + \lambda^A_i + \lambda^B_j + \lambda^C_k + \beta^{AB}\,u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + I\,(i = j) + I\,(i = k) + I\,(j = k) + I\,(i = j = k) \quad (4)$$

$$3.\log(m_{ijk}) = \lambda + \lambda^A_i + \lambda^B_j + \lambda^C_k + \beta^{AB}\,u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + \beta^{ABC}u_iv_jw_k \quad (5)$$

$$4.\log(m_{ijk}) = \lambda + \lambda^A_i + \lambda^B_j + \lambda^C_k + \beta^{AB}\,u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + I\,(i = j) + I\,(i = k) + I\,(j = k) \quad (6)$$

$$5.\log(m_{ijk}) = \lambda + \lambda^A_i + \lambda^B_j + \lambda^C_k + \beta^{AB}\,u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + I\,(i = j = k) \quad (7)$$

$$6.\log(m_{ijk}) = \lambda + \lambda^A_i + \lambda^B_j + \lambda^C_k + \beta^{AB}\,u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + \beta^{ABC}u_iv_jw_k + I\,(i = j) + I\,(i = k) + I\,(j = k) \quad (8)$$

$$7.\log(m_{ijk}) = \lambda + \lambda^A_i + \lambda^B_j + \lambda^C_k + \beta^{AB}\,u_iv_j + \beta^{AC}u_iw_k + \beta^{BC}v_jw_k + \beta^{ABC}u_iv_jw_k + I\,(i = j) + I\,(i = k) + I\,(j = k) + I\,(i = j = k) \quad (9)$$

The explanations of the terms in the models are given below:

m_{ijk}, is the expected frequency for cell, (i, j, and k), which is calculated over the model. λ reflects the constant term.

λ^A_i reflects the effect of ith decision of rater A. λ^B_j reflects the effects of jth decision of rater B. λ^C_k reflects the effect of kth decision of rater C while i, j, k = 1, ..., R. .

$$\sum_{i=l}^{R} \lambda_i^A = \sum_{i=l}^{R} \lambda_j^B = \sum_{i=l}^{R} \lambda_{ik}^C = 0.$$

β^{AB}, β^{BC}, β^{AC} are association parameters between 2 raters. However, β^{ABC} is the association parameter between 3 raters. u_i, v_j, and w_k are respectively the score values that belong to raters 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.

I (i = j), I (i = k), and I (j = k) are agreement parameters between 2 raters. However,

I (i = j = k) is the agreement parameter between 3 raters.

Pathological sample diagnosis

The application of the models introduced in the second section will be given on a real data set. According to Agresti, 7 pathologists classified each of 118 slides in whether there is carcinoma in the uterine cervix, based on the most involved lesions (9) such as:

1) Negative
2) Atypical squamous hyperplasia
3) Carcinoma in situ
4) Squamous carcinoma with early stromal invasion
5) Invasive carcinoma.

Since in the original study as many sampling zero cells were encountered, the categories (3), (4), and (5) are combined like in Perkins and Becker (10); therefore, the number of levels was reduced. Only the results belonging to the pathologists A, B, and C were used. The results corresponding to 3 pathologists are given in Table 1.

Table 1. The results of 118 slides according to 3 raters.

A	**B**	**C**		
		1	**2**	**3**
1	**1**	18	4	0
	2	1	1	0
	3	0	2	0
2	**1**	2	3	0
	2	3	4	0
	3	4	10	0
3	**1**	0	0	0
	2	0	2	1
	3	3	16	44

For this evaluation, the models defined in the second section are solved and the results are given in Table 2.

Among the models analyzed, the fitting to other models except Model 1 was found significant. Since there is more than one fitted model, to find the best model the Akaike Information Criterion ($AIC = G^2 - 2df$) was calculated. The model with a minimum AIC is the best one. According to this rule, Model 5 is the best model.

Kappa coefficient for multiraters given in Shoukri (11) was found as $\kappa = 0.4839$. According to this kappa's value, there is a moderate agreement between the raters.

Based on the results in Table 3, the highest relation is between pathologists A-C and the lowest relation is between pathologists B and C.

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

Models	**Likelihood Ratio Statistics**	**df**	**P value**	**AIC**
1	45.994	16	0.000	-
2	14.567	13	0.335	−11.433
3	19.679	16	0.235	−12.321
4	17.227	14	0.244	−10.773
5	15.936	16	0.457	−16.064
6	15.990	13	0.250	−10.010
7	14.155	12	0.291	−9.845

Table 3. Parameter estimations and odds ratios for Model 5.

Parameter	**Estimation**	**St. Error**	**Z**	**Odds ratio**
β^{AB}	1.390	0.391	3.551*	4.015
β^{AC}	1.273	0.438	2.905*	3.571
β^{BC}	0.331	0.339	0.975	1.392
I (i = j = k)	0.885	0.417	2.121*	2.423

*P < 0.05

The above Model 5 is expressed in terms of the conditional local log odds ratios for

$i, j, k = 1, \ldots, R - 1$.

$$
\begin{aligned}
\log\theta_{ij(k)} &= \beta^{AB} + I\,(i = j = k) && i = j = k \quad \text{or} \quad i + 1 = j + 1 = k,\\
&= \beta^{AB} - I\,(i = j = k) && i < j = k \quad \text{or} \quad j < i = k,\\
&= \beta^{AB} && \text{otherwise.}
\end{aligned}
$$

$$
\begin{aligned}
\log\theta_{i(j)k} &= \beta^{AC} + I\,(i = j = k) && i = j = k \quad \text{or} \quad i + 1 = j + 1 = k,\\
&= \beta^{AC} - I\,(i = j = k) && i < k = j \quad \text{or} \quad k < i = j,\\
&= \beta^{AC} && \text{otherwise.}
\end{aligned}
$$

According to the odds ratios in Table 3, probability of giving i + 1 decision rather than i of pathologist A is 4 times higher than giving i + 1 decision rather than i of pathologist B. The probability of giving the same decision of each of the 3 pathologists is exp(0.885) = 2.423 times higher than giving a different decision.

Discussion

The agreement of decisions between 2 or more raters is measured by kappa coefficient if the scale is nominal. If the scale is ordinal then weighted kappa coefficient is used for agreement. Raters to be statistically independent from each other, requires kappa coefficient to be zero. However, it does not mean that the raters must be independent. In addition, as kappa is an agreement index, it does not give an agreement between categories. For example, researcher can deal with in which cases raters agree with each other and in which cases they do not. In studies where the scale is ordinal, weighted kappa coefficient is used rather than kappa coefficient. There are different score definitions for the score values. The choice of those scores will affect the weighted kappa's value. However, as kappa coefficient is a single value it does not give any possibility of detailed interpretations. When the scale is ordinal, there is no possibility to interpret the progressive decisions of raters with kappa coefficient. Therefore, especially in recent years, rather than calculating kappa, the analysis of agreement with log-linear models has become widespread (6,7).

Log-linear models help cross tables to be interpreted with odds ratios. In studies where the scale is nominal, log-linear models are used for the agreement only. In studies where the scale is ordinal, the use of log-linear models for agreement with association together to differentiate association from agreement is important. In these kinds of model equations, association parameters and agreement parameters are estimated separately. Local odds ratios, calculated with the help of estimated parameters, help in the interpretation of cross tables. These models can be solved with 'general log-linear' statistical software.

In this study modeling was carried out for an agreement between more than 2 raters. In the second section, 7 different models were introduced that investigate agreement and association separately and together.

References

1. Cohen JA. Coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960; 20: 37-46.
2. Agresti A. Categorical data analysis. New York: Wiley; 2002.
3. Fleiss J, Cohen J, Everitt, BS. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin 1969; 72: 323-7.
4. 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.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74.

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

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

8. Lawal, B. Categorical data analysis with SAS and SPSS applications. Mahwah, New Jersey, London: Lawrence Erlbaum Associates Publishers; 2003.

9. Agresti A. Loglinear modeling of pairwise interobserver agreement on a categorical scale. Statistics in Medicine 1992; 11: 101-14.

10. Perkins SM., Becker MP. Assessing rater agreement using marginal association models. Statistics in Medicine 2002; 21: 1743-60.

11. Shoukri M. Measures of interobserver agreement. Chapman & Hall/CRC, Boca Raton, FL, USA; 2004.