Original Article

Assessing the inter-rater reliability for nominal, categorical and ordinal data in medical sciences

Bizhan Shabankhani

Assistant Professor, Department of Biostatistics, Health Sciences Research Center, Faculty of Public Health, Mazandaran University of Medical Sciences, Sari, Iran.

Abstract

Background: Inter-rater Reliability (IRR) measures the reliability of raters. The number of raters, the type of variable evaluated, and some other constraints have a direct impact on the use of this tool. Failure to pay attention to these points will increase the error and will harm the results of the research. **Materials and Methods:** The present study is an analytical one. The samples of this research were randomly selected from students and interns of faculties and hospitals of Mazandaran University of Medical Sciences. The samples of this research were selected from Sari Faculty of Health and Sari Accident and Burn hospital. **Results:** This study shows how the agreement between the two raters is calculated concerning a nominal variable with and without the factor of chance. Besides, by understanding the amount of agreement between the two raters in the presence of an ordinal variable, we will become acquainted with the calculation of the agreement between more than two raters. **Discussion:** Some limitations, such as the quantitative and qualitative variables, the number of raters, the existence of missing data directly affects the size of the IRR; according to these points, the appropriate IRR for data to measure the agreement between raters is used.

Keywords: Inter-rater reliability, Cohen's kappa, Weighted kappa, Fleiss' kappa, Icc

Introduction

Validity and reliability are the two main concepts to evaluate instruments in a study. Reliability is related to random errors and validity is related to regular errors therefore it is possible to reduce the random error by increasing the sample size. It is also possible to reduce the regular error using more precise tools [1, 2].

Reliabilities with the three categories of tools, raters, and research samples are measured separately by the relevant indicators. The Inter-Rater Reliability Index (IRR) measures the reliability of raters. In this paper, the rater is a term used to describe people who rank people in the study, such as a trained research assistant who ranks people [1]. Diagnosing radiological images or diagnosing diseases based on expert judgment are excellent examples of that. Reliability among raters is not reported in many medical studies. A lack of attention to this indicator can harm the results of a study [1, 3]. The key to work with this indicator is to be aware of its types, as well as the conditions for choosing the best type of IRR for survey data.

Materials and Methods

The present study is an analytical one. The samples of this research were randomly selected from students and interns of faculties and hospitals of Mazandaran University of Medical Sciences. The samples of this research were selected from Sari Faculty of Health and Sari Accident and Burn Hospital.

Results

Percent Agreement

The percent agreement can measure the amount of agreement between two or more raters. For example, if there is agreement among the raters in 7 samples out of 10, the percent agreement will be 70% [2, 3]. Table 1 below shows information about three raters (students) and 6 samples (teachers). Each teacher is judged by the students (raters) and scores from one to six. The first column shows the number of the teacher, the next three columns show the score of the raters to each teacher, and the next three columns show the agreement of the three raters in pairs. The last column also shows the agreement ratio between the raters for each sample.

Address for correspondence: Bizhan Shabankhani. Assistant Professor, Department of Biostatistics, Health Sciences Research Center, Faculty of Public Health, Mazandaran University of Medical Sciences, Sari, Iran.

How to cite this article: Shabankhani, B. Assessing the inter-rater reliability for nominal, categorical and ordinal data in medical sciences. Arch Pharma Pract 2020;11(S4):144-8.

Table 1: Information from Three Raters and 6 Samples to Determine the Percent Agreement

Units	Rater A	Rater B	Rater C	A/B	A/C	B/C	Agreement
1	3	3	2	1	0	0	1.3
2	5	5	5	1	1	1	3.3
3	4	4	2	1	0	0	1.3
4	6	1	4	0	0	0	0.3
5	2	2	5	1	0	0	1.3
6	3	3	3	1	1	1	3.3

Now we calculate the mean agreement of the samples (the last column).

$$\frac{\frac{1}{3}+\frac{3}{3}+\frac{1}{3}+\frac{0}{3}+\frac{1}{3}+\frac{3}{3}}{6} = 0.5(\%50)$$

The percent agreement among the three raters is 50% in this case [3].

Two important points about the percent agreement reduce the use of this indicator in research. The first point is the number of raters. If the number of raters increases, the number of possible pairs of raters will also increase, which makes the estimated percentage of the agreement greater than the actual value of the report given that part of this agreement is due to chance [4, 5].

Cohen's Kappa

The kappa coefficient is the most widely used indicator in measuring IRR. It depends on the data, therefore, there are several types of kappa coefficient. Choosing the appropriate kappa coefficient based on the data set is very important. There are prerequisites for calculating the kappa coefficient that must be verified before calculating.

1- Raters should be independent.
2- The scale of the rater's judgment should be clear.
3- Raters should make a judgment of the same observations.
4- The number of levels for raters' judgments should be equal.
5- The consistency of the raters must also be determined. If raters are consistent, we use Cohen's kappa, if raters are not consistent and are randomly selected, we use Fleiss' kappa.
6- It is assumed by the null hypothesis that the value of the observed agreement of the raters is equal to the value of the chance agreement [4, 6, 7].

In the first step, we introduce Cohen's kappa (Kc) coefficient, which measures the agreement between the two raters. The variable studied, in this case, is also a nominal variable of two or more states. In a study to identify three different races of one animal species, two evaluators were used. The results of that study are given after evaluation in Table 2 below, during which 36 animals were used.

Table 2: Race Diagnosis of Three Samples of an Animal Species by Two Raters

Rater2 \ Rater1	Race1	Race2	Race3	Total
Race1	**9**	3	1	13
Race2	4	**8**	2	14
Race3	2	1	**6**	9
Total	15	12	9	36

The two raters agree on only 23 of the diagnoses, some of which may have been based on chance. That's why we use the Kc statistic to measure the actual agreement of two raters. To calculate Kc, we must know the maximum number of agreements (which is the total number of people present in the study) in addition to the number of observed agreements and the number of expected agreements. The expected number of agreements is calculated as follows and only for the cells in which the agreement takes place.

$$\frac{13*15}{36}+\frac{14*12}{36}+\frac{9*9}{36} = 5.42+4.67+2.25 = 12.34$$

$$K_c = \frac{\text{number of observed agreements} - \text{number of expected agreements}}{\text{maximum number of agreements} - \text{number of expected agreements}} = \frac{23-12.34}{36-12.34} = 0.45$$

A value of less than 0.7 is usually not allowed to accept the agreement between two raters. In this example, raters measured a nominal variable, but raters may have to measure an ordinal variable [8]

Weighted Kappa

Another type of kappa coefficient we are introduced to in this section is the weighted kappa (Kw), during which we encounter two raters and an ordinal variable. There are two types of Kw (linear - quadratic) [9-12]. In the linear model, the distance between different levels is considered the same, for example, in an ordinal variable with 5 different weight levels in a linear set, equal distances are considered (1, 0.75, 0.5, 0.25, and 0) while in the quadratic model, distances are considered different (1, 0.937, 0.750, 0.437 and 0). In this section, we will learn how to calculate the weighted kappa in an example. Table 3 shows the results of pain diagnosis in 100 burn patients measured by two experts.

Table 3: Information from Two Experts and 100 Samples to Calculate the Weighted Kappa

Doctor A \ Doctor B	Painless	Mild Pain	Moderate Pain	Severe Pain	Total
Painless	15	3	1	1	20
Mild Pain	4	18	3	2	27
Moderate Pain	4	5	16	4	29
Severe Pain	1	2	4	17	24
Total	24	28	24	24	100

To calculate the weight, we consider the distance between different levels to be the same. This means that the distance between the 1st floor and the 2nd floor will be equal to the distance between the 2nd floor and the 3rd floor. Accordingly, we form Table 4 of distances.

Table 4: Distance Table

Doctor A \ Doctor B	Painless	Mild Pain	Moderate Pain	Severe Pain
Painless	0	1	2	3
Mild Pain	1	0	1	2
Moderate Pain	2	1	0	1
Severe Pain	3	2	1	0

After determining the distance between the levels, we obtain the weight of each cell. The weight of each cell is calculated in two linear and quadratic models based on the following formulas.

$$\text{Cell weight in linear model } = 1 - \frac{|\text{Distance}|}{\text{Max possible distance}}$$

$$\text{Cell weight in the quadratic model } = 1 - \frac{(\text{Distance})^2}{(\text{ Max possible Distance })^2}$$

Table 5 shows the weights calculated in the two linear and quadratic models.

Table 5: Weights in Linear and Quadratic Models

Doctor A \ Doctor B		Painless	Mild Pain	Moderate Pain	Severe Pain
Painless Mild Pain	Linear	1	0.67	0.33	0
	Quadratic	1	0.89	0.56	0
Moderate Pain	Linear	0.67	1	0.67	0.33
	Quadratic	0.89	1	0.89	0.56
Painless Mild Pain	Linear	0.33	0.67	1	0.67
	Quadratic	0.56	0.89	1	0.89
Moderate Pain	Linear	0	0.33	0.67	1
	Quadratic	0	0.56	0.89	1

Table 6 shows the values shown in Table 3 by percentage.

Table 6: Values Observed in Table 3 by Percentage

Doctor A \ Doctor B	Painless	Mild Pain	Moderate Pain	Severe Pain
Painless	0.15	0.03	0.01	0.01
Mild Pain	0.04	0.18	0.03	0.02
Moderate Pain	0.04	0.05	0.16	0.04
Severe Pain	0.01	0.02	0.04	0.17

Table 7 also shows the expected values by percentage.

Table 7: Expected Values in Percentage

Doctor A \ Doctor B	Painless	Mild Pain	Moderate Pain	Severe Pain
Painless	0.048	0.056	0.048	0.048
Mild Pain	0.0648	0.0756	0.0648	0.0648
Moderate Pain	0.0696	0.0812	0.0696	0.0696
Severe Pain	0.0576	0.0672	0.0576	0.0576

To calculate the weighted kappa coefficient in the linear model, we proceed as follows. First, we multiply the percentage of observed values in each cell by the weight of the same cell in the linear model. Then we sum all the obtained values together to get $P_{observed}$:

$P_{observed}=0.8538$

Then we do the same for the Table of expected values and weights in the linear model to get $P_{observed}$:

$P_{expected}=0.597$

Now, we calculate the weighted kappa coefficient in the linear model using the following formula.

$$K_{lw} = \frac{P_{observed} - P_{expected}}{1 - P_{expected}} = \frac{0.8538 - 0.597}{1 - 0.597} = 0.648$$

The calculation of the weighted kappa coefficient in the quadratic model (KQW) is similar to that method only instead of linearly weighted Tables, we use quadratically weighted Tables [4].

Mathematical Forms: The coefficients presented above make it possible to calculate this statistic for only two raters. This mathematical form is generalized for calculating the IRR for

three or more raters. Fleiss has done this generalization using the average Kappa value calculated. This index is used for Likert, response packet, nominal, categorical, and ordinal data [13-15]

We will provide an example of how to calculate the Fleiss' Kappa in this section. Twelve psychology residents were used to diagnose the severity of the disease in 6 patients. The severity of the disease is divided into 5 levels. To calculate the agreement, we first draw a Table in which the number of columns is the number of levels (K = 5) and the number of rows is the number of samples (N = 6) present in the study (Table 8). The numbers in column ij are also the number of evaluators assigned to class j for example i. For the first line, all raters rated the severity of the disease in the first sample as the fifth level.

Table 8: Information Table of 6 Samples and 12 Psychologists to Calculate Fleiss' Kappa

N_{ij}	None	Mild	Moderate	Severe	Very Severe	P_i
1	0	0	0	0	12	1.000
2	0	1	5	4	2	0.258
3	0	0	3	5	4	0.288
4	0	3	7	2	0	0.379
5	2	2	7	1	0	0.348
6	7	5	0	0	0	0.470
Total	9	11	24	12	18	
P_j	0.125	0.153	0.333	0.167	0.250	

In this Table, P_j is the ratio of the sum of all the observations in level j to the maximum score that can be obtained at the same level. The maximum score for each level is N * n = 6 * 12 = 72 points.

$$p_1 = \frac{9}{72} = 0.125$$

Other P_j values are calculated similarly. But we calculate P_i according to the following command.

$$p_1 = \frac{1}{12(12-1)}(0^2 + 0^2 + 0^2 + 0^2 + 12^2 - 12) = 1.000$$

Then we calculate the average of P_i values that show the observed values.

$$\overline{p_{observed}} = \frac{1 + 0.258 + \cdots + 0.470}{6} = 0.457$$

The expected value ($\overline{p_{\text{expected}}}$) in this case, is calculated as follows.

$$\overline{p_{\text{expected}}} = 0.125^2 + 0.153^2 + 0.333^2 + 0.167^2 + 0.250^2 = 0.240$$

Fleiss' Kappa coefficient is calculated based on the following formula [16]

$$K_f = \frac{0.457 - 0.240}{1 - 0.240} = 0.285$$

The mathematical form for calculating the amount of agreement is the same in all cases, and the way the components are calculated differs only according to the state of the data.

Discussion

This study was designed to investigate and get acquainted with the amount of agreement between raters and their measurement coefficients. Initially, this study showed how the agreement between two raters on a nominal variable can be calculated. Then we learned how to calculate this amount without the intervention of the luck factor, using the kappa statistic. In the next step, we are introduced to the weighted kappa, which calculates the amount of agreement on an ordinal variable between two raters. It was also shown that if the number of raters increases, Fleiss' kappa can be used to measure the agreement between the raters. Despite being introduced to these coefficients in IRR measurement, there are still limitations in this area that the proposed coefficients cannot calculate the IRR value correctly if these limitations occur. One of those limitations is the presence of quantitative data, in which case the amount of agreement between the raters is measured using the ICC index and the Bland-Altman index.

Limitations on the number of raters, limitations on the number of categories, impossibility of calculating the IRR in the presence of incomplete or missing information, the impossibility of calculating the IRR for large or small samples, are some of the other things that can be mentioned. The Krippendorff's alpha index is another tool to measure the agreement among raters, with the difference that it has much fewer limitations than other coefficients [16-18]. This coefficient can measure the amount of agreement between the raters despite all the mentioned limitations. One of the reasons that Krippendorff's alpha Index is a more reliable index than other indices is that it calculates the difference between the raters instead of calculating the amount of agreement. Due to the computational complexity of this coefficient, the details related to it have not been mentioned in this article. It is hoped that in the future readers will be more familiar with this coefficient [6, 19].

Conclusion

Most errors in calculating reliability are due to the number of raters and measurement scales of the relevant variable. Failure to pay attention to this point will lead to unrealistic reliability calculations. In this article, we have addressed the

point for **Nominal, Categorical and Ordinal Data**. For other cases, you can refer to the other articles of the authors.

References

1. Mohammadbeigi A, Mohammadsalehi N, Aligol M. Validity and reliability of the instruments and types of measurments in health applied researches. Journal of rafsanjan university of medical sciences. 2015;13(12):1153-70.
2. Yawn BP, Wollan P. Interrater reliability: completing the methods description in medical records review studies. American Journal of Epidemiology. 2005;161(10):974-7.
3. Kottner J. The difference between reliability and agreement. journal of clinical Epidemiology. 2011;64:701-2.
4. Nurjannah I, Siwi SM. Guidelines for analysis on measuring interrater reliability of nursing outcome classification. Int J Res Med Sci. 2017;5(4):1169-75.
5. Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960;20(1):37-46.
6. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in quantitative methods for psychology. 2012;8(1):23.
7. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica. 2012;22(3):276-82.
8. Sullivan AD. Determining an inter-rater agreement metric for researchers evaluating student pathways in problem solving Iowa State University 2014.
9. Vanbelle S. A new interpretation of the weighted kappa coefficients. Psychometrika. 2016;81(2):399-410.
10. Warrens MJ. Weighted Kappas for 3×3 Tables. Journal of Probability and Statistics. 2013;2013:1-9.
11. Warrens MJ. Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables. Statistical Methodology. 2011;8(2):268-72.
12. Yang Z, Zhou M. Weighted kappa statistic for clustered matched-pair ordinal data. Computational Statistics & Data Analysis. 2015;82(c):1-18.
13. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy. 2005;85(3):257-68.
14. Beyer WH. CRC Standard Mathematical Tables. 31st ed: Boca Raton, Fla, CRC Press; 2002.
15. Kotz S, Balakrishnan N, Read CB, Vidakovic B. Encyclopedia of Statistical Sciences. : Wiley; 2006.
16. Fleiss J. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76(5):378-82.
17. Antonia Z, Stefanie C, Lars M, Andre K. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? BMC Medical Research Methodology 2016;16(93):1-10.
18. Shabankhani B, Yazdani Charati J, Shabankhani K, Kaviani Cherati S. Survey of agreement between raters for nominal data using krippendorff's Alpha. Archives of Pharmacy Practice 2020;11(1):160-4.
19. Dooley K. questionnaire programming language. interrater reliability report 2017.