Original Article

Survey of agreement between raters for nominal data using krippendorff's Alpha

Bizhan Shabankhani[1,2]*, Jamshid Yazdani Charati[1,2], Keihan Shabankhani[3], Saeid Kaviani Cherati[4]

[1] Department of Biostatistics, School of Health, Mazandaran University of Medical Sciences, Sari, Iran. [2] Health Sciences Research Center, Addiction Research Institutes, Mazandaran University of Medical Sciences, Sari, Iran. [3] Department of Medicine, School of Medicine, Mazandaran University of Medical Sciences, Sari, Iran.[4] Department of Public Health, School of Health, Mazandaran University of Medical Sciences, Sari, Iran.

Abstract

Background: Most of the indicators used to measure IRR have limitations such as the number of raters, the number of categories, the type of variable (nominal, ordinal, interval, ratio), and missing data. The krippendorff's Alpha coefficient is the only indicator among the IRR indices, which, despite all the limitations, calculates the agreement among the raters with the appropriate confidence. **Materials and Methods:** The study used indexed articles in Google scholar databases, Medline, Scopus, Springer, ScienceDirect, published in English since 2000 and also the collected data from the Design projects of sound and vibration control systems in the industry that has been done in Sari faculty of public Health. **Results:** In this paper, we will introduce the method of calculating binary and nominal data in the presence of two or more raters, and in both cases the existence or absence of missing data. **Conclusion:** Most of the coefficients used to measure the agreement between raters will not provide a satisfactory reliability if there are some limitations. The krippendorff's Alpha statistics can be used as an efficient statistic in assessing the extent of agreement between evaluators replacing other statistics. Of course, it should be noted that the calculations of this index are more complex than other indicators, but provide a higher reliability.

Keywords: krippendorff's Alpha, missing data, nominal data, inter-rater reliability

INTRODUCTION

The basis of decision making is based on the values measured by the relevant experts in many applied sciences such as health sciences, social sciences, and technical sciences, etc. For this reason, the accuracy of the measurements is very important. One of the important points about the measured values is the person who performed the measurement. The term “Reliability” is used in the researches where the measurements were made. For example we expect to see the same size when measuring a person's height during a day. Seeing different sizes is a justification for using Reliability. Reliability of measuring instruments, research subjects and raters can be measured. There are different ways to measure the reliability of each of the above three. These indices are generally calculated as a coefficient and we will have a more reliable index by increasing them. [1-5]

Inter-rater Reliability (IRR) is measured by a set of coefficients. Most of the coefficients used to measure IRR have limitations such as the number of raters, the number of categories, the type of variable (nominal, ordinal, interval, ratio) and missing data.

Krippendorff’s alpha coefficient is an efficient instrument for assessing reliability among raters. The krippendorff's Alpha coefficient is the only indicator among the IRR indices, which, despite all the limitations, calculates the agreement among the raters. [6, 7]

The number 1 indicates complete agreement and the number zero indicates the least amount of agreement among the raters in this index. Typically for $\alpha \geq 0.823$ a good agreement is considered, acceptable at $0.667 \leq \alpha \leq 0.823$ and unacceptable at $\alpha < 0.667$.

Based on the limitations presented and the analysis used, there are four methods for calculating the Krippendorff’s alpha.

Address for correspondence: Mr. Bizhan Shabankhani. Assistant Professor, Department of Biostatistics, Health Sciences Research Center, Faculty of Public Health, Mazandaran University of Medical Sciences, Sari, Iran.
E-Mail: shabankhani@yahoo.com

How to cite this article: Shabankhani, B., Yazdani Charati, J., Shabankhani, K., Kaviani Cherati, S. Survey of agreement between raters for nominal data using krippendorff's Alpha. Arch Pharma Pract 2020;10(S1):160-4.

1- Binary data, two raters, no missing data
2- Nominal data, two raters, no missing data
3- Nominal data, any number of raters, with missing data
4- Any data, any number of raters, with missing data

We describe the first three methods and we consider the fourth stage separately because of the workload in this study.

Materials and Methods

This study is a kind of review study conducted in 2018. The study used indexed articles in Google scholar databases, Medline, Scopus, Springer, ScienceDirect, published in English since 2000 and also the collected data from the Design projects of sound and vibration control systems in the industry. Missing data, Krippendorff's alpha, Inter-rater Reliability, Nominal data.

Results

In this section, we get to know the Krippendorff's alpha calculations in three modes. For further understanding each section is followed by an example.

a) Calculation of Krippendorff's alpha for binary data, two raters, no missing data

Two students were asked to comment on the usefulness of general courses offered at the university. In the first step, we build the Reliability data Matrix based on the raters and units under investigation.

Table 1. General Course Poll Information

courses students	1	2	3	4	5	6	7	8
First student	0	1	0	0	0	0	0	1
Second student	0	1	0	1	1	0	0	1

In the second step, we form Coincidences Matrix based on this matrix. Coincidences Matrix is a square matrix that it's rows and columns hold the Diagnostic variable values (the values of 1 and 0 in this example). The elements within this matrix are pairs formed by the recognition of two raters , due to the values given in this example, it is possible to see four pairs (0 and 0) and (0 and 1) and (1 and 0) and (1 and 1). It should be noted that two pairs can be formed for each unit. For example, for the first unit two pairs (1 and 0) and (0 and 1) can be formed. In fact, the Coincidences Matrix shows the number of pairs formed [8].

	0	**1**	
0	8	2	**10**
1	2	4	**6**
	10	6	**16**

Number of (0,0) pairs $O_{00}=8$

Units 1,3,6,7 each form two pairs (0 and 0)

Number of (1,1) pairs $O_{11}=4$

Number of (0,1) pairs $O_{01}=2$

Number of (1,0) pairs $O_{10}=2$

The number of times that the diagnostic variable detected zero by two raters, in the total 16 times. $n_0 = 10$

The number of times that the diagnostic variable detected one by two raters, in the total 16 times. $n_1 = 6$

The total number of pairs formed n=2N

The elements of the main diameter indicates the extent of agreement and the Elements that are symmetrically on both sides of the main diameter shows the extent of disagreement between raters.

$$\propto = 1 - \frac{\text{observed disagreement}}{\text{expected disagreement}} = 1 - (n-1)\frac{o_{01}}{n_1 * n_2}$$

$$\propto = 1 - (16-1)\frac{2}{10 * 6} = 0.5$$

b) Calculation of Krippendorff's Alpha for Nominal data, two raters, No missing data

Two physicians asked to prioritize 10 patients referred to the emergency department. The two doctors categorized patients into three categories. After detection, we form the reliability data matrix.

Table 2. Prioritizing Emergency Patients Information

Patients raters	1	2	3	4	5	6	7	8	9	10
First physician	A	B	B	C	A	B	C	C	A	B
Second physician	A	A	B	B	A	B	B	C	A	A

Based on this matrix, we form the Coincidences Matrix.

$$\begin{array}{c|ccc|c} & A & B & C & \\ \hline A & 6 & 2 & 0 & 8 \\ B & 2 & 4 & 2 & 8 \\ C & 0 & 2 & 2 & 4 \\ \hline & 8 & 8 & 4 & 20 \end{array}$$

As stated earlier, the elements of the main diameter indicates the extent of agreement and the Elements that are symmetrically on both sides of the main diameter shows the extent of disagreement between raters.

$$n_1 = 8\ , n_2 = 8\ , n_3 = 4\ \ , n = 2N = 20$$

The value of Krippendorff's alpha in this case is calculated using the following formula [8, 9]:

$$\propto = 1 - \frac{\text{observed disagreement}}{\text{expected disagreement}} = \frac{(n-1)\sum_i o_{ii} - \sum_i n_i(n_i - 1)}{n(n-1) - \sum_i n_i(n_i - 1)}$$

$$\propto = \frac{(20-1)(6+4+2) - (8*7+8*7+4*3)}{20*(20-1) - (8*7+8*7+4*3)}$$

$$\propto = 0.406$$

c) Calculation of Krippendorff's alpha for nominal data, any number of raters , with missing data

We asked 4 students to identify sound at six points in a vocational training center in one study. We see below the reliability data matrix of this study. m_u is the number of diagnoses recorded for the unit U in this table. For example, only three raters stated their opinion for the first unit and one rater has not commented on what we consider to be missing therefore $m_1 = 3$.

Table 3. Six-point sound measurement information at a vocational training center

Points Raters	1	2	3	4	5	6
First student	0	2	3	1	2	**3**
Second student	1	2	2	1	2	**0**
Third student	1	2	2	1	0	**3**
Fourth student	1	3	2	1	2	**3**
m_u	3	4	4	4	3	**3**

The elements of Coincidences Matrix are then calculated as in the previous states except that after we obtain each element from the matrix (o_{ij}), we divide it into m_u-1.

$$O_{ij} = \frac{\text{Number of } i,j \text{ pairs in unit u}}{m_i - 1}$$

Accordingly, the Coincidences Matrix will be visible as we can see below.

$$\begin{array}{c|ccc|c} & 1 & 2 & 3 & \\ \hline 1 & 9 & 0 & 0 & 9 \\ 2 & 0 & 6 & 2 & 8 \\ 3 & 0 & 2 & 3 & 5 \\ \hline & 9 & 8 & 5 & 22 \end{array}$$

The calculation of Krippendorff's alpha in this case is based on the previous formula [8, 10]

$$\propto = 1 - \frac{\text{observed disagreement}}{\text{expected disagreement}} = \frac{(n-1)\sum_i o_{ii} - \sum_i n_i(n_i - 1)}{n(n-1) - \sum_i n_i(n_i - 1)}$$

$$\propto = \frac{(22-1)(9+6+3) - (9*8+8*7+5*4)}{22*(22-1) - (9*8+8*7+5*4)}$$

$$\propto = 0.79$$

Discussion

Choosing the appropriate method to assess Inter-raters Reliability requires awareness of the limitations of the research. These limitations include the number of raters, the type of data, the missing data, the accuracy required (For example, if the wrong decision is made, there will be no danger to public health or financial cost) etc. Most of the coefficients used to measure agreement among raters do not provide a good reliability despite these limitations. [11-13]

The Krippendorff's alpha coefficient is the only index among the IRR indices that calculates the extent of agreement between the raters with acceptable reliability despite all limitations. Table 4 clearly shows the strengths of this index compared to other IRR indices [14, 15].

Percentage of agreement is widely used in such cases as the first indicator. But this index has the least flexibility, considering only nominal data and It's not useful in the presence of missing data. Also, there is no fixed and specific value among the IRR indices for delineating the amount of acceptable agreement [14, 16, 17].

Table 4. Comparison of IRR indices in presence of research limitations

IRR	Data	Missing Data	Number of Raters	The effect of 'chance' in agreement is minimized?	**General agreement on the significance of a numeric result?**
Percent Agreement	Nominal	No	2	No	**No**
Bennett et asks	Nominal	No	2	No *	**No**
Scott's p_i	Nominal	No	2≥	No *	**No**
Fleiss's Kappa	Nominal	No	2≥	No *	**No**
Cohen's Kappa	Nominal	No	2≥	No *	**No**
Gwet	Nominal	Yes	2≥	No *	**No**
Weighted Kappa	Nominal	No	2	No *	**No**
Krippendorff's Alpha	All Data	Yes	2≥	Yes	**Yes** **

** Krippendorff's Alpha considers 0.823 as the cut point.

Cohen's Kappa is another controversial statistic to be very cautious about when using it. Prevalence, odds, raters independence, and impact on diagnosis and some other factors can strongly influence the results of Kappa statistics. Cohen's Kappa statistic provides acceptable reliability when conditions are met but If these conditions are not met, the Kappa statistics are severely affected therefore there is a widespread disagreement over the use of Kappa statistics for final evaluation. Researchers find the use of kappa statistics more appropriate to assess the independence of raters except where they are confident of the correctness of the requirements [14].

Scott's p_i statistic is a coefficient for nominal data with two raters that calculates agreement despite some limitations but When the data is nominal and we have two evaluators, the Krippendorff's alpha formula changes to the Scott's p_i formula [8, 14].

Also, Fleiss's Kappa statistic, which is the generalized form of Cohen's kappa to more than two raters, exhibits some incompatibility in obtaining observation ratios by pair counting in the small samples. In this case, Krippendorff's alpha can be used instead of Fleiss's Kappa. Interestingly, the Krippendorff's Alpha is rarely used in place of Fleiss's Kappa, especially in medical researches While Krippendorff has widely discussed the advantages and disadvantages of this work [17-21]. The ICC is in fact a special case of Krippendorff's alpha as a coefficient of agreement [14]. As you can see, the Krippendorff's alpha statistic can be an efficient statistic to replace other statistics in measuring the extent of agreement between raters. It should be noted that the computation of this index is more complex than the other indices, but it offers higher reliability, especially in cases where there are no perfect conditions for research [14].

References

1. Hayes AF, Krippendorff K. describe and provide SPSS and SAS macros for computing oalpha, its confidence limits and the probability of failing to reach a chosen minimum, 2007.(http://www.comm/ohioi-state.edu/ahayes/SPSS%20programs0kalpha.htm).
2. Reference manual of the irr package containing the kripp. alpha() function (http://cran.r-project.org/web/packages/irr/irr.pdf#page.16) for the platform-independent statistics package R.
3. the alpha resources page. (http://cswww.essex.ac.uk/Research/nle/arrau/alpha.htm).
4. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. communicatiion methods and measure, 2007; 1, .77-89.
5. Krippendorff K. content analysis: An introduction to its methodology, 3rd edition. Thousand Oaks, CA:Sage, 2013.
6. Brennan RL. An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement. 2001;38(4):295-317.
7. Krippendorff K. content analysis: An introduction to its methodology, Chapter 11, 2rd edition. Thousand Oaks, CA:Sage publications. 2004; 211-56 p.
8. Krippendorff K. Computing Krippendorff's alpha-reliability. 2011.
9. Freelon DG. worked examples for nominal intercoder Reliability 2009. Available from: http://dfreelon.org/recal/recal-worked-examples.pdf.
10. MC 6110 Computing Intercoder Reliability. Available from: http://drkblake.com/wp-content/uploads/2015/03/Computing-Krippendorff-Alpha.pdf.
11. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in quantitative methods for psychology. 2012;8(1):23.
12. Kappa Coefficients: A Critical Appraisal. Available from: http://www.john-uebersax.com/stat/kappa.htm.

13. Warrens MJ. Five ways to look at Cohen's kappa. Journal of Psychology & Psychotherapy. 2015;5(4):1.
14. Nili A, Tate M, Barros A. A critical analysis of inter-coder reliability methods in information systems research. 2017.
15. Xie Q. Agree or disagree? A demonstration of an alternative statistic to Cohen's kappa for measuring the extent and reliability of agreement between observers. Unpublished manuscript. 2013.
16. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of clinical epidemiology. 1990;43(6):543-9.
17. Fleiss JL, Nee JC, Landis JR. Large sample variance of kappa in the case of different sets of raters. Psychological bulletin. 1979;86(5):974.
18. LePage R, Billard L. Exploring the limits of bootstrap. New york: division of biostatistics, standford university: John Wiley & Sons; 1992.
19. MacPherson P, Choko AT, Webb EL, Thindwa D, Squire SB, Sambakunsi R, et al. Development and validation of a global positioning system–based "Map Book" system for categorizing cluster residency status of community members living in high-density urban slums in Blantyre, Malawi. American journal of epidemiology. 2013;177(10):1143-7.
20. Devine A, Taylor SJ, Spencer A, Diaz-Ordaz K, Eldridge S, Underwood M. The agreement between proxy and self-completed EQ-5D for care home residents was better for index scores than individual domains. Journal of clinical epidemiology. 2014;67(9):1035-43.
21. Krippendorff K. Commentary: A dissenting view on so-called paradoxes of reliability coefficients. Annals of the International Communication Association. 2013;36(1):481-99.