South Eur J Orthod Dentofac Res

Short Communication

SEJODR

StaTips Part III: Assessment of the repeatability and rater agreement for nominal and ordinal data

***Perinetti, Giuseppe** **

** Department of Medical, Surgical and Health Sciences, University of Trieste, Italy*

ABSTRACT

Assessment of the repeatability and rater agreement for nominal and ordinal data is an important procedure in medical research. This analysis applies to several cases including: i) replicate recordings by the same rater using the same parameter; ii) agreement between/among raters using the same parameter; and iii) agreement between/among different parameters recorded by the same rater. Herein, the most common coefficients of agreement are presented along with criteria for their proper selection.

Perinetti G. StaTips Part III: Assessment of the repeatability and rater agreement for nominal and ordinal data. South Eur J Orthod Dentofac Res. 2017;4(1):3-4.

FRAMING OF THE PROBLEM

In the last StaTips Part II the problem of the assessment of the method error, i.e. repeatability of the measurements, has been introduced. [1] While in the last paper the case of continuous parameters has been taken into consideration, the corresponding procedures for nominal and ordinal data sets are presented herein. It is important to point out that repeatability for nominal (either dichotomous or not) and ordinal data sets may include the following cases: i) replicate recordings by the same rater using the same parameter; ii) agreement between/among raters using the same parameter; and iii) agreement between/among different parameters recorded by the same rater.

COEFFICIENTS OF REPEATABILITY/AGREEMENT

Among the methods of estimating agreement for nominal and ordinal data sets, the Cohen's kappa, [2] Fleiss' kappa [3] and Kendall's W [4] are among the most used, and the correct choice of the coefficient depends on the different situation under analysis (Table 1). All of these coefficients range from zero for no repeatability/agreement to 1 for perfect repeatability/agreement, and the following standards for strength of repeatability/agreement for the coefficient have proposed: 0.01-0.20, slight; 0.21-0.40, fair; 0.41-0.60, moderate; 0.61-0.80, substantial; and >0.80 almost perfect. [5] Irrespective of the used coefficient, the reporting of the confidence interval of each estimation is also recommended. Several statistic packages may calculate these coefficients including SPSS software (SPSS® Inc., Chicago, Illinois, USA) and MedCalc® software (MedCalc Software, Mariakerke, Belgium).

COHEN'S KAPPA

The simplest and most common cases that may be encountered in research are those regarding the intra-rater agreement assessing the presence/absence of a given condition between time points, or between-rater agreement assessing the presence/absence of a given condition at the same time, or when the same rater has to assess the presence/absence of a given condition using two different scoring systems. In all of these cases, the unweighted Cohen's kappa is the coefficient of choice. Of note, the unweighted Cohen's kappa may be used for nominal data, either dichotomous (i.e. presence/absence of a condition) or not. The latter case is less frequent in orthodontics and it applies when the rater has to assess the presence/absence of several mutually exclusive categories.

Corresponding Author:
Perinetti Giuseppe
Struttura Complessa di Clinica Odontoiatrica e Stomatologica,
Ospedale Maggiore,
Piazza Ospitale 1, 34129 Trieste, Italy
e-mail: G.Perinetti@fmc.units.it

A slightly different situation comes when the parameter that is being used has an ordinal nature, which means that its values represent a quantity on a (non-linear) scale. [6] Therefore, the outcome 1 would be less than the outcome 2, which in turn would be less than the outcome 3, and so on. The skeletal maturational stages are an example of ordinal scale. In such a case, it would be best to consider not only the absolute agreement between raters (or between parameters), but also the relative agreement that depends on how far the repeated recordings lies along the ordinal scale. Therefore, the Cohen's kappa may be subjected to a weighting procedure, which may be either linear or quadratic. [7] Usually, the linear weighted Cohen's kappa is preferred over the quadratic weighted one, as the former procedure makes the coefficient less sensitive to the number of categories, i.e. stages in the ordinal scale. [7]

FLEISS' KAPPA AND KENDALL'S W

A further situation not considered yet is that of more than two raters which are involved in an analysis of agreement using a nominal (either dichotomous or not) parameter. For such a case, an extension of the Cohen's kappa, referred to as the Fleiss' kappa is to be used. While the Cohen's kappa may be unweighted or weighted, the Fleiss' kappa is always unweighted.

Finally, a last (and less frequent) situation is that when more than two raters are involved in an analysis of agreement using a parameter of ordinal nature. In such a case, the Kendall's W, also referred to as Kendall's coefficient of concordance, has to be applied.

Table 1. *Most common statistical methods to assess the repeatability and rater agreement for nominal and ordinal data according to the number of raters/parameters*

Number of raters/parameters	Nature of the parameter	
	Nominal	**Ordinal**
Two raters, same parameter Two parameters, same rater	Unweighted Cohen's kappa	Weighted Cohen's kappa
More than two raters, same parameter More than two parameters, same rater	Fleiss' kappa	Kendall's W

REFERENCES

1. Perinetti G. StaTips Part II: Assessment of the repeatability of measurements for continuous data. South Eur J Orthod Dentofac Res. 2016;3(2):33-34.
2. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46.
3. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions, 3rd ed. Wiley. 2003.
4. Kendall MG, Babington Smith B. The Problem of m Rankings. Ann Math Statist. 1939;10:275-87.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
6. Perinetti G. StaTips Part I: Choosing statistical test when dealing with differences. South Eur J Orthod Dentofac Res. 2016;3(1):4-5.
7. Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology. 1996;7(2):199-202.