Roel Popping

Introduction to Interrater Agreement for Nominal Data

EXTRAS ONLINE

Springer

Introduction to Interrater Agreement for Nominal Data

Roel Popping

Introduction to Interrater Agreement for Nominal Data

Springer

Roel Popping
Department of Sociology
University of Groningen
Groningen, The Netherlands

Additional material to this book can be downloaded from http://extras.springer.com.

ISBN 978-3-030-11670-5 ISBN 978-3-030-11671-2 (eBook)
https://doi.org/10.1007/978-3-030-11671-2

Library of Congress Control Number: 2018967726

Mathematics Subject Classification (2010): 62, 97, 60

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Marika.

Preface

This book was born out of the feeling of need to help other researchers at times when I am not around. They must be able to set up, carry out, and report research themselves in which agreement between raters plays an important role. This concerns the extent to which there is consensus between independent raters in the classification of a sample of research units. It is not important here whether it concerns encoding texts or answers to open questions in surveys, observing in a behavioral study, diagnosing by an insurance physician, or determining PTSD by a psychiatrist. The book does not contain a detailed description of any of these types of research, but contains adequate material for a researcher to be able to build up such research.

The book contains the most important agreement measures we have and connects them with the most common research situations in which these measures are used. This should be sufficient for most researchers to start with. More important, the book tries to give backgrounds, due to which the researcher is expected to gain more understanding about establishing agreement between raters.

My thanks go to all those with whom I have collaborated in this kind of research in one way or another. They came to me with their problems. But I have learned a lot from those problems. This is reflected in the book.

Groningen, The Netherlands
December 2018

Roel Popping

Contents

Part I Theory and Practice

1 Introduction . . . 3
References . . . 8

2 Reliability and Validity . . . 11
2.1 Reliability . . . 11
2.2 Validity . . . 16
References . . . 19

3 Interrater Agreement . . . 21
3.1 Units, Categories, Raters . . . 24
3.1.1 Units . . . 25
3.1.2 Categories . . . 33
3.1.3 Raters . . . 39
3.2 Requirements to Be Fulfilled by Agreement Indices . . . 43
3.3 Chance Expected Agreement . . . 52
3.3.1 Chance Expected Agreement Based on Uniform Marginal Probability . . . 52
3.3.2 Chance Expected Agreement Based on Marginal Homogeneity with Replacement . . . 53
3.3.3 Chance Expected Agreement Based on Marginal Homogeneity Without Replacement . . . 55
3.3.4 Chance Expected Agreement Based on Marginal Heterogeneity . . . 56
3.3.5 Chance Expected Agreement Discussion . . . 58
3.4 How to Proceed When Agreement Is Satisfactory? . . . 63
3.5 A Number of Relevant Research Situations . . . 65
3.6 Different Views on Agreement . . . 67

3.7 Selection and Training of Raters 69
3.8 Interval or Ordinal Level of Measurement 72
References 75

Part II Formulas

4 Indices 81
4.1 A Posteriori Method of Coding 81
4.1.1 Two Raters 82
4.1.2 Weights 85
4.1.3 Intracategory Agreement for Two Raters 87
4.1.4 Example with More Than Two Raters 92
4.1.5 More Than Two Raters 93
4.1.6 Intracategory Agreement for More Than Two Raters 99
4.1.7 Elaboration Research Situations 103
4.1.8 Simultaneous and Majority Agreement 117
4.1.9 Supposed Problems Due to Chance Correction 119
4.1.10 Test on Homogeneous Use of Categories 125
4.2 A Priori Method of Coding 126
4.3 Miscellaneous 131
4.3.1 Inference: Standard Errors, Significance, Interval Estimation 131
4.3.2 Standards for Assessing and Reporting Reliability 136
4.3.3 Computer Programs 137
4.4 Indices that Do Not Fulfill the Requirements 137
References 142

Index 147

Notation

In the text, the following symbols are used:

N	Number of units
m	Number of raters
c	Number of categories
r	Number of categories by the other rater in case this rater has an own set of categories
$f_{ij(gh)}$	Number of units in cell [i, j] of the agreement table for raters g and h. In case there are only two raters, this might be written as f_{ij}
$p_{ij(gh)}$	Proportion of units in cell [i, j] of the agreement table for raters g and h
$f_{(s)ij}$	Number of units in cell [i, j] of the agreement table for unit s based on the assignments by two raters
$f_{i(g)}$	Marginal total in row i of the agreement table, marginal total assigned by rater g to category i
$p_{i(g)}$	Marginal proportion in row i of the agreement table, marginal total assigned by rater g to category i
p_i	Proportion of units assigned to category i over all raters
$p_{i(mG)}$	Proportion of units assigned to category i over all raters from group G
$p_{i(mG+H)}$	Proportion of units assigned to category i over all raters from groups G and H together
w_{ij}	Weight in cell [i, j] of the matrix of weights
n_{si}	Number of times unit s has been assigned to category i by all raters or in all ratings
n_i	Number of times a unit has been assigned to category i by all raters or in all ratings
y_{sig}	Unit s has been coded in category i by rater g (y = 1) or not (y = 0)
v_{sii}	Number of identical codings assigned by two raters to unit s in the situation of multicoding

$v_{si(g)}$ Number of different codings used by rater g to classify unit s in the situation of multicoding

$f_{i_1 i_2 \ldots i_m}$ Number of agreeing assignments by all m raters to category i

$p_{i_1 i_2 \ldots i_m}$ Proportion of agreeing assignments by all m raters to category i

The symbol # is used to indicate marginal totals in tables.

Part I
Theory and Practice

Chapter 1
Introduction

When a physician examines a patient, it is desirable that the findings do not change when this patient is examined by a different physician. The patient will not feel confident if physicians seriously differ in opinion. Of course, this does not only apply to physicians, but applies to those who make judgments in general, especially in situations where it is impossible to establish the truth in an objective way.

In many fields of science, research units like persons (patients), events, texts, or broadcastings, observations are classified with respect to some characteristic into one of a limited set of categories, groups of units regarded as having particular shared characteristics. Such a classification is necessary because this is the only way to get a classification of the research units. A physician, for example, diagnoses patients or recognizes symptoms with respect to the type of disease, and a personnel officer classifies jobs with respect to the type of skills required. A rater in a text analysis study categorizes an article in a newspaper as mainly dealing with home or foreign news, but might also code open-ended interview data into analyzable terms. In observation studies, one looks at types of behavior that are exposed like, for example, the showing of solidarity or of antagonism. Even astronomers might classify star formations as a spiral, beam, or one of the other types that have been distinguished among these formations. In specific investigations, pictures, sounds, and gestures might be judged. Dozens of examples from many fields could be added. In many situations, the measurement involves at least some ambiguity in the characteristics of interest in the unit. The measurement task often involves a subjective moment when the decision is to be made which category applies most for a specific unit. The person, rater, responsible for the coding is supposed to know how to deal with this ambiguity. This rater is an expert or had a training in how to perform the coding. The only way to find out whether the decision by the rater can be trusted is by having the classification also done by another rater, who is equally skilled as the first one and who operates independently from that first one. The scores should not represent the rater's personnel opinion, but should be in line with the view that is common in the culture in which the study is performed. The more identical the assignments per unit, the more trust an investigator can have in the final classifications. This situation is often found in the fields of the social,

R. Popping, *Introduction to Interrater Agreement for Nominal Data*,
https://doi.org/10.1007/978-3-030-11671-2_1

behavioral, political and communication sciences, marketing, biology, and medicine. The variable that is constructed in this way will be confronted with other variables in the study.

In any scientific research, associations between different may be blurred by errors in the assessment of characteristics. The possibility exists that important associations are overlooked. Associations that are discovered may be hard to interpret because different studies may result in different quantitative conclusions. The specification of disagreement may be a first step to a better agreement among raters: Which disagreements do frequently occur and which raters often have different opinions. This may lead to a further standardization of the way of judging.

Interrater agreement is the widely used term for the extent to which independent raters evaluate a characteristic of a message or artifact and reach the same conclusion. The classifications that are made always concern one specific characteristic. The importance of interrater agreement lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured. This book deals with the question of how to compute interrater agreement. More precisely, it will be on agreement on data measured at a nominal level of measurement. At this level, one simply names or categorizes units. As an example, in case we would look at the variable 'eye color,' the categories might be 'blue,' 'green,' 'brown,' and 'hazel.' In text analysis studies, one has to think about whether a sentence in a text is about politics, economics, or none of these, as in the earlier example. In observation studies, one looks at whether behavior that is shown expresses solidarity, antagonism, or something else. A psychiatrist classifies patients as schizophrenics, manic depressives, psychotic depressives, or in some other psychotic group. The essential point about nominal scales is that they do not imply any ordering among the categories. A nominal scale consists of a (limited) number of mutually exclusive and exhaustive categories. Higher levels of measurement are the ordinal level, where the categories also are ordered, and the interval level, which is one more step higher. The interval level of measurement not only classifies and orders the measurements, but it also specifies that the distances between any two succeeding categories is the same. A consequence of the restriction to data on the nominal level is that agreement statistics based on Pearson product moment correlations or upon analysis of variance techniques (which assume interval categories) are ruled out.

If the scores assigned to units by two or more raters are (almost) the same based on some criterion, the coding task is said to have a good reliability. This indicates most of all that there seems to be little measurement error and that the codings are reproducible. Some researchers, as we will come to see, go further. They add that the raters basically are exchangeable and that such an exchange does not affect the coding process. Having reliable data means the investigator has generated data in his or her study that can be relied upon and therefore can be studied, theorized, or used in pursuit of practical decisions or further analysis.

Table 1.1 contains the assignments by two raters in a text analysis study using the theoretical premise that people motivate themselves and others by persuading them that their actions are possible, impossible, inevitable, or contingent. This persuasion is accomplished using reality claims (Popping and Roberts 2009).

Table 1.1 Two raters' assignments of four reality claims

		Rater 2					
		Possible	Impossible	Inevitable	Contingent	Not coded	#
Rater 1	Possible	217	24	0	0	15	256
	Impossible	2	120	0	0	2	124
	Inevitable	0	0	84	0	2	86
	Contingent	0	1	0	8	0	9
	Not coded	4	3	3	2	0	12
	#	223	148	87	10	19	487

Sentences containing such claims had to be coded into one of these four categories. At first glance, the categories seem to be difficult to understand. After a careful explanation however, the rater can recognize them very well. Abstract categories like these are often used when judgments are made. It is possible that one rater noted a claim while the other did not. Therefore, an additional category—not coded—was necessary. An example of a sentence to be coded is the following:

> An ethical point of view is usually justified in case of all human actions, because its subjects, humans *cannot* be separated from their actions (Népszabadság, April 1, 1994, p. 14).

This sentence does not contain a reality claim that motivates a person; the word 'cannot' as used here does not refer to an intention. Another example is the sentence:

> Hungarian society is just learning the rules of democracy, and developing the rules of the game, which *must* be adopted by politicians (Népszabadság, October 15, 1993, p. 15).

This sentence is encoded as a reality claim, even though its grammatical subject is not a person. This is because its use of 'must' as the reality claim *inevitable* is in passive voice and has people (namely politicians) as its semantic subject. Coded is the rater's interpretation of the sentence, usually it is not known which choice is the correct one, but, if there are clear rules to follow, raters should come to the same assignment. The raters had the same training. They are equally skilled, and therefore, they should be interchangeable. It is during the coding process that data are generated for the investigation. Only after data have been generated, the investigator is able to link each individual datum and the whole data set to the units of interest.

The comparison of all assignments on which the raters agreed with respect to the type of reality claim results in the score on an agreement index, which is a descriptive index. The simplest index, but not the best as will be explained, consists of the proportion of units on which both raters agree. This is the proportion of units on the diagonal of the table showing the assignments by the raters; for the data in Table 1.1, the proportion of observed agreement, P_o, is (217 + 120 + 84 + 8 + 0)/ 487 = 0.88. Disagreement is found with respect to the units that are not on the diagonal of the table.

The difference between reliability and agreement is described very well: ‘Interrater reliability provides an indication of the extent to which the variance in the ratings is attributable to differences among the rated subject. … Interrater agreement represents the extent to which the different judges tend to assign exactly the same rating to each object’ (Tinsley and Weiss 2000: 98). In the above study, the question was whether the codes are a good reflection of reality, that is, to say whether they are valid. If the degree of agreement is higher than a certain criterion, it is stated that the codes are reliable. This is an important step toward validity. By relating the variable with these codes as scores to variables that relate to something else from the content of the sentences from which they are taken, for example, their rationale (politics, economics, and so on) and at the time (year) on which they were published, it is possible to get view on developments in a culture regarding the direction in which that culture should change according to the residents. Most investigators connect agreement to reliability as is done here.

The example presented above concerns data at the nominal level of measurement. In practice, the rater sometimes needs two steps to come to a decision. First, the rater must decide whether the attribute is at present or not. If it is at present the second decision is how it shows itself. The first decision with respect to the example used is whether a reality claim is found in the sentence under investigation or not. If so, the next decision concerns the type of reality claim. The two decisions are always reported in one table and one outcome of the index is reported. But actually, there are two sources of disagreement. Raters may differ in the *definition of the attribute itself* (does the sentence contain a reality claim: yes or no); or in their definitions of *specific categories* (which type of reality claim is at present in the sentence).

The number of different interrater agreement indices found in the literature is remarkably high. This is partly due to the discovery that the measurement of agreement is not only relevant in the simple situation in which two raters have each classified a sample of N units ($N > 1$) into the same set of c pre-specified categories. Many variations, extensions, and refinements of this simple situation are met in empirical research, and moreover, the ways and goals of measuring agreement exhibit a similar variation. Such a specific setting is called here an ‘empirical situation.’ In this text, many of these situations will be discussed. In practice, we frequently see that (a variant of) a measure is used in a situation where it should not be used. As we will see later, this often concerns the kappa-statistic, a measure which is very popular. This book will not give a complete overview of the literature and will also not treat all specific research situations that might be distinguished; for developments up to 1990, see Popping (1992). Since books have appeared covering (almost) all at that moment available measures, for example: Gwet (2014), Shoukri. (2011).

The idea in this book text is to help the investigator, who is to use interrater agreement indices, get started. We start from a sample of units that is at least rated twice.

Agreement for data at an ordinal or interval level of measurement is not considered other than that it is mentioned that these levels can be approached by using weights. Usually, the reliability for data at these levels is computed in another way.

However, one can use weights in the formulas to cover these levels. Weights will have to be entered into the indices for these data. The issue will come back shortly at the end of Chap. 3.

Studying agreement is not useful in qualitative research. The concept of reliability is considerably overrated in this type of research. Here, coders often notice different things in the data or interpret them differently. If they do, then these issues need to be resolved. The idea of 'saturation' is relevant here. Campbell et al. (2013) have tried to develop a method for computing agreement in a qualitative study. They found that especially the way transcripts are unitized biases coding and inflates agreement scores and therefore interrater reliability. However, one must look carefully at what is meant by qualitative research. It occurs often that studies deal with 'qualitative research,' but that these are actually descriptive (quantitative) studies. This can be a study where the investigator is still looking for the best set of categories to be used. It is possible that units are available, and a category is assigned to each unit. However, the categories belong to a set that is developed by the rater, and each rater uses an own set. Here too, interrater agreement can be computed (Popping 1983a). Afterward one can see whether and where these sets of categories differ and, if so, find out what the rater has been looking for. This might help in defining the set of categories to be used in the actual study. In a real qualitative study however, where the goal is to construct theories, one does not work with well-defined units.

The judgments of only one unit by many raters are not a problem of interrater agreement. Such data one gets when one asks lots of people (by preference a sample) to judge, for example, a painting on some (one-dimensional) scale. The score denotes an average appreciation for the unit. For such, the reader might start with Lindell and Brandt (1997).

Also, a note on terminology is necessary. The name 'rater' denotes the agent that produces a classification (although in some applications this is not one human being). In many texts, the words 'coder,' 'judge,' 'observer,' 'annotator,' or 'classifier' are used, sometimes even the description 'diagnostic test.' In general, 'units' is used for that what is classified (although these may be human beings and not things as in our example of patients). In literature at least also, the words 'respondent,' 'person,' 'patient,' 'observation,' 'subject,' 'object,' 'phenomenon,' and 'unit of analysis' are found. To make it more complex, often not the complete unit is investigated, but only an attribute of the unit, i.e., a quality or characteristic that is a particular part of a unit. Where possible I will try to use the word 'unit,' but at some places 'person' or 'observation' will be used as otherwise the text sounds curious.

One might distinguish between 'sampling units' and 'recording units.' The first are the entities that are sampled and that are usually used in the statistical analysis. The second refers to the part that is recorded. Sometimes the two coincide, but often a sampling unit contains several recording units. In the study from which Table 1.1 was taken, the sampling units are newspaper editorials that were sampled and the recording units were all sentences in these editorials that contain a reality claim. Some editorials do not contain such a claim at all; others contain several claims. In observation studies, the test person is generally the sampling unit; the occurrence of