Springer Series on Epidemiology and Public Health

Suhail A.R. Doi
Gail M. Williams *Editors*

Methods of Clinical Epidemiology

Springer

Springer Series on Epidemiology and Public Health

For further volumes:
http://www.springer.com/series/7251

Suhail A.R. Doi • Gail M. Williams
Editors

Methods of Clinical Epidemiology

Springer

Editors
Suhail A.R. Doi
Clinical Epidemiology Unit
School of Population Health
University of Queensland
Herston, Queensland
Australia

Gail M. Williams
Department of Epidemiology & Biostatistics
School of Population Health
University of Queensland
Herston, Queensland
Australia

ISSN 1869-7933 ISSN 1869-7941 (electronic)
ISBN 978-3-642-37130-1 ISBN 978-3-642-37131-8 (eBook)
DOI 10.1007/978-3-642-37131-8
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013940612

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book was written to fill the gap that exists in the methods of epidemiology of interest to clinical researchers. It will enable a reader who is currently undertaking research to get key information regarding methodology. It will also help health care personnel from all fields (doctors, nurses, allied health, dentists, pharmacists, etc.) to obtain an effective understanding of methodology useful to research in their field as we cover the unique methods not covered properly in current research methods texts. The classic theoretical focus is avoided because we believe that research must be based on understanding guided by the reader's knowledge of the methodology.

Part I begins by introducing readers to the methods used in clinical agreement studies. It is written to suit beginners but without turning off intermediate users. Qualitative and quantitative agreement are presented, and this section explains how we can utilize these methods and their strengths and weaknesses. Part II shows readers how they can interpret and conceptualize diagnostic test methodologies and ends with an introduction to diagnostic meta-analyses. Part III takes the reader through methods of regression for the binomial family as well as survival analysis and Cox regression. Here, the focus is on methods of use to clinical researchers. These methods have different names and multiple interpretations, which are explained. It is important to know what associations you are interested in and know what data are available and in what form they can be used. An in-depth discussion of the use of these methods is presented with a view to giving the reader a clear understanding of their utility and interpretation. Part IV deals with systematic reviews and meta-analyses. A step-by-step approach is used to guide readers through the key principles that must be understood before undertaking a meta-analysis, with particular emphasis on newer methods for bias adjustment in meta-analysis, an area in which we have considerable expertise.

We thank Lorna O'Brien from authorserv.com for her dedicated help with the editing of this book and Federica Corradi Dell'Acqua, Editorial Manager for Biomathematics & Statistics at Springer for continuous advice throughout the publication process. Finally, we realize that this first edition may include inconsistencies and mistakes, and we welcome any suggestion from readers to improve its content.

30 June 2012 Suhail A.R. Doi
Brisbane Gail M. Williams

Acknowledgments

Every effort has been made to trace rights holders, but if any have been inadvertently overlooked the publishers would be pleased to make the necessary arrangements at the first opportunity.

List of Abbreviations

ADA	Adenosine deaminase activity
AGME	Accreditation Council for Graduate Medical Education
ANA	Antinuclear antibodies
ANOVA	Analysis of variance
AUC	Area under the curve
BAK	Bias-adjusted kappa
BF	Body fat
BMI	Body mass index
BP	Blood pressure
CADTH	Canadian Agency for Drugs and Technology in Health
CASP	Critical Appraisal Skills Programme
CEBM	Centre for Evidence Based Medicine
CF	Correction factor
CHF	Congestive heart failure
CI	Confidence interval
CL	Confidence limits
CT	Computed tomography
CV	Coefficient of variation
DOR	Diagnostic odds ratio
DXA	Dual-energy X-ray absorptiometry
ECT	Electroconvulsive therapy
ELISA	Enzyme-linked immunosorbent assay
EPHPP	Effective Public Health Practice Project
ES	Effect size
ESR	Erythrocyte sedimentation rate
ESS	Effective sample size
FN	False-negative
FP	False-positive
FPR	False-positive rate

GLM	Generalized Linear Model
HIV	Human immunodeficiency virus
HSROC	Hierarchical summary receiver operator characteristic
HTA	Health technology assessment
ICC	Intraclass correlation coefficient
ICC	Intraclass correlation coefficient
IID	Independent and identically distributed
ITT	Intention-to-treat
LAG	Lymphangiography
LCL	Lower confidence limit
LR	Likelihood ratio
MERGE	Method for Evaluating Research and Guideline Evidence
MeSH	Medical Subject Heading
MI	Myocardial infarction
MRI	Magnetic resonance imaging
MSE	Mean squared error
NLM	National Library of Medicine
NOS	Newcastle-Ottawa Scale
NPV	Negative predictive value
NSS	Numerical sum score
NTP	Negative test probability
OCLC	Online Computer Library Center
OFIA	Operational financial impact assessment
PABAK	Prevalence-adjusted-bias-adjusted kappa
PICO	Population, intervention or exposure, comparison, outcome
PPV	Positive predictive value
PT	Pertussis toxin
QCR	Qualitative rating on level of components
QE	Quality effect
QOR	Qualitative overall rating
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RCT	Randomized controlled trial
RE	Random effect
REVC	Random effects variance component
RLR	Ratio of the likelihood ratio
ROC	Receiver operating characteristic
ROM	Range of motion
RR	Relative risks
SD	Standard deviation
SE	Standard error
SEM	Standard error of the measurement
SLE	Systemic lupus erythematosus
SMD	Standardized mean difference
TcB	Transcutaneous bilirubin

TN	True-negative
TP	True-positive
TPR	True-positive rate
TSB	Total serum bilirubin
UCL	Upper confidence limit
WHO	World Health Organization

Contents

Contributors

Víctor Abraira Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal, CIBER en Epidemiología y Salud Pública (CIBERESP) and Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain

Cristian Baicus Clinical Epidemiology Unit, Bucharest, and Associate Professor of Internal Medicine and Carol Davila University of Medicine and Pharmacy, Bucharest, Romania

Jan J. Barendregt University of Queensland, School of Population Health, Brisbane, Australia

Justin Clark University of Queensland, Brisbane, Australia

Suhail A.R. Doi University of Queensland, School of Population Health, Brisbane, Australia and Princess Alexandra Hospital, Brisbane, Australia

Maren Dreier Hannover Medical School, Institute for Epidemiology, Social Medicine and Health Systems Research, Hannover, Germany

Rosmin Esmail Knowledge Translation, Research Portfolio, Alberta Health Services, Calgary, AB, Canada

Abhaya Indrayan Department of Biostatistics and Medical Informatics, University College of Medical Sciences, Delhi, India

Rajeev Kumar Malhotra Department of Biostatistics and Medical Informatics, University College of Medical Sciences, New Delhi, India

Adedayo A. Onitilo Marshfield Clinic - Weston Center, Weston, USA and University of Wisconsin Medical School, Madison, USA

María Nieves Plana Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal, CIBER en Epidemiología y Salud Pública (CIBERESP) and Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain

Orit Shechtman Department of Occupational Therapy, College of Public Health and Health Professions, University of Florida, Gainesville, FL, USA

Sophie Vanbelle Department of Methodology and Statistics, School of Public Health and Primary Care, Maastricht University, Maastricht, The Netherlands

Robert Ware University of Queensland, School of Population Health, Brisbane, Australia

Gail M. Williams University of Queensland, School of Population Health, Brisbane, Australia

Javier Zamora Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal, CIBER en Epidemiología y Salud Pública (CIBERESP) and Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain

Part I
Clinical Agreement

Chapter 1
Clinical Agreement in Qualitative Measurements

The Kappa Coefficient in Clinical Research

Sophie Vanbelle

Abstract Agreement between raters on a categorical scale is not only a subject of scientific research but also a problem frequently encountered in practice. For example, in psychiatry, the mental illness of a subject may be judged as light, moderate or severe. Inter- and intra-rater agreement is a prerequisite for the scale to be implemented in routine use. Agreement studies are therefore crucial in health, medicine and life sciences. They provide information about the amount of error inherent to any diagnostic, score or measurement (e.g. disease diagnostic or implementation quality of health promotion interventions). The kappa-like coefficients (intraclass kappa, Cohen's kappa and weighted kappa), usually used to assess agreement between or within raters on a categorical scale, are reviewed in this chapter with emphasis on the interpretation and the properties of these coefficients.

Introduction

The problem of rater agreement on a categorical scale originally emerged in human sciences, where measurements are traditionally made on a nominal or ordinal scale rather than on a continuum. For example, in psychiatry, the mental illness of a subject may be judged as light, moderate or severe. Clearly two psychiatrists assessing the mental state of a series of patients do not necessarily give the same grading for each patient but we would expect some agreement between them (inter-rater agreement). In the same way, we could observe some variation in the assessment of the same patients by a psychiatrist on two occasions (intra-rater agreement). Agreement studies, therefore, became crucial in health, medicine and

S. Vanbelle (✉)
Department of Methodology and Statistics, School of Public Health and Primary Care, Maastricht University, Maastricht, The Netherlands
e-mail: Sophie.vanbelle@Maastrichtuniversity.nl

S.A.R. Doi and G.M. Williams (eds.), *Methods of Clinical Epidemiology*, Springer Series on Epidemiology and Public Health,
DOI 10.1007/978-3-642-37131-8_1,

life sciences. They provide information about the amount of error inherent to any diagnosis, score or measurement.

Agreement has to be distinguished from the concept of reliability. When elements (objects, subjects, patients, items) are evaluated by two raters (observers, judges, methods), agreement refers to the degree of closeness between the two assessments within an element (i.e. classification of each element in the same category by the two raters). By contrast, reliability refers to the degree of differentiation between the elements (i.e. the two raters give the same relative ordering of the elements). Good reliability is essential when the purpose is to assess the correlation with other measures (e.g. severity of the mental illness and autonomy) because of the well-known attenuation effect. Good agreement is, on the other hand, imperative for clinical decision making (e.g. prescribing a treatment for a specific patient based on the seriousness of the mental illness). Two kinds of agreement are usually distinguished. Inter-rater agreement refers to a sample of elements assessed with the same instrument by different raters; the term intra-rater agreement is used when a sample of elements is assessed on two occasions by the same rater using the same instrument.

Several coefficients for quantifying the agreement between two raters on a categorical scale have been introduced over the years. Cohen's (1960) kappa coefficient is the most salient and the most widely used in the scientific literature. Cohen (1968) extended the kappa coefficient to weighted kappa coefficients to allow for some more important disagreements than others (e.g. disagreements between the categories light and severe may be viewed as more important than between light and moderate). Kraemer (1979) defined the intraclass kappa coefficient by assuming that the two raters have the same marginal probability distribution. All these coefficients belong to the kappa-like family and possess the same characteristic: they account for the occurrence of agreement due to chance only.

An example used through this chapter to illustrate the use and the computation of the various kappa coefficients is presented in the next section. The third section focuses on binary scales. Kappa coefficients are introduced for nominal and ordinal scales in the fourth and fifth sections, respectively. Then, before drawing conclusions, the distinction between the concepts of agreement and association is illustrated on an example.

Example

Cervical ectopy, defined as the presence of endocervical-type columnar epithelium on the portio surface of the cervix, has been identified as a possible risk factor for heterosexual transmission of human immunodeficiency virus. Methods for measuring the cervical ectopy size with precision are therefore needed. Gilmour et al. (1997) conducted a study to compare the agreement obtained between medical raters by direct visual assessment and a new computerized planimetry method. Photographs of the cervix of 85 women without cervical disease were assessed for cervical ectopy by three medical raters who used both

Table 1.1 4×4 contingency table resulting from the direct visual assessment of cervical ectopy size by two medical raters on 85 women in terms of frequency

	Medical rater 2				
Medical rater 1	Minimal	Moderate	Large	Excessive	Total
Minimal	13	2	0	0	15
Moderate	10	16	3	0	29
Large	3	7	3	0	13
Excessive	1	4	12	11	28
Total	27	29	18	11	85

assessment methods. The response of interest, cervical ectopy size, was an ordinal variable with $K = 4$ categories: (1) minimal, (2) moderate, (3) large and (4) excessive. The classification of the 85 women by two of the three raters via direct visual assessment is summarized in Table 1.1 in terms of frequency. We will determine the agreement between these two raters on each category separately and on the four-point scale.

Binary Scale

The simplest case is to determine the agreement between two raters who have to classify a sample of N elements (subjects, patients or objects) into two exhaustive and mutually exclusive categories (e.g. diseased/non-diseased). For example, women can be classified as having (1) or not having (0) an excessive ectopy size. The observations made by the two raters can be summarized in a 2×2 contingency table (Table 1.2), where n_{jk} is the number of elements classified in category j by rater 1 and category k by rater 2, $n_{j.}$ the number of elements classified in category j by rater 1 and $n_{.k}$ the number of elements classified in category k by rater 2. By dividing these numbers by the total number of observations N, the corresponding proportions p_{jk}, $p_{j.}$, $p_{.k}$ are obtained $(j, k = 1, 2)$. The proportions $p_{1.}$ and $p_{2.}$ determine the marginal distribution of rater 1 and $p_{.1}$ and $p_{.2}$ the marginal distribution of rater 2. The marginal distribution refers to the distribution of the classification of one rater, irrespective of the other rater's classification.

Cohen's Kappa Coefficient

Intuitively, it seems obvious to use the sum of the diagonal elements in Table 1.2 to quantify the level of agreement between the two raters. It is the proportion of elements classified in the same category by the two raters. This simplest agreement index, usually denoted by p_o, is called the *observed proportion of agreement*,

$$p_o = \frac{n_{11} + n_{22}}{N} = p_{11} + p_{22}.$$

Table 1.2 2×2 contingency table corresponding to the classification of N elements on a binary scale by two raters in terms of frequency (left) and proportion (right)

	Rater 2				Rater 2		
Rater 1	1	0	Total	Rater 1	1	0	Total
1	n_{11}	n_{12}	$n_{1.}$	1	p_{11}	p_{12}	$p_{1.}$
0	n_{21}	n_{22}	$n_{2.}$	0	p_{21}	p_{22}	$p_{2.}$
Total	$n_{.1}$	$n_{.2}$	N	Total	$p_{.1}$	$p_{.2}$	1

However, this coefficient does not account for the fact that a number of agreements between the two raters can occur purely by chance. If the two raters randomly assign the elements on a binary scale (e.g. based on the results of a tossed coin), the proportion of agreement between them is only attributable to chance. Therefore, Cohen (1960) introduced the *proportion of agreement expected by chance* as

$$p_{\mathrm{e}} = \frac{n_{1.}n_{.1} + n_{2.}n_{.2}}{N^2} = p_{1.}p_{.1} + p_{2.}p_{.2}$$

It is the proportion of agreement expected if rater 1 classifies the elements randomly with a marginal distribution $(p_{.1}, p_{.2})$ and rater 2 with a marginal distribution $(p_{1.}, p_{2.})$. Cohen corrected the observed proportion of agreement for the proportion of agreement expected by chance and scaled the result to obtain a value 1 when agreement is perfect (all observations fall in the diagonal cells of the contingency table), a value 0 when agreement is only to be expected by chance and negative values when the observed proportion of agreement is lower than the proportion of agreement expected by chance (with a minimum value of -1). Specifically, Cohen's kappa coefficient is written as

$$\hat{\kappa} = \frac{p_{\mathrm{o}} - p_{\mathrm{e}}}{1 - p_{\mathrm{e}}}. \tag{1.1}$$

Cohen's kappa coefficient is more often used to quantify inter-rater agreement than intra-rater agreement because it does not penalize the level of agreement for differences in the marginal distribution of the two raters (i.e. when $p_{1.} \neq p_{.1}$). Different marginal distributions are expected in the presence of two raters with different work experience, background or using different methods.

Intraclass Kappa Coefficient

The intraclass kappa coefficient was derived by analogy to the intraclass correlation coefficient for continuous outcomes and is based on the common correlation model. This model assumes that the classifications made by the two raters are interchangeable. In other words, the two raters are supposed to have the same marginal probability distribution (i.e. the probability of classifying an element in category

1 is the same for the two raters). This is typical of a test–retest situation where there is no reason for the marginal probabilities to change between the two measurement occasions. The resulting index is algebraically equivalent to Scott's index of agreement and can be viewed as a special case of Cohen's kappa coefficient.

The *observed proportion of agreement* is the same as in the case of Cohen's kappa coefficient

$$p_{\mathrm{oI}} = \frac{n_{11} + n_{22}}{N} = p_{11} + p_{22} = p_{\mathrm{o}}$$

but the proportion of agreement expected by chance is determined by

$$p_{\mathrm{eI}} = \bar{p}_1^2 + (1 - \bar{p}_1)^2$$

where $\bar{p}_1$ estimates the probability, common to the two raters, of classifying an element in category 1, namely $\bar{p}_1 = (p_{1.} + p_{.1})/2$. The intraclass kappa coefficient is then defined by

$$\hat{\kappa}_{\mathrm{I}} = \frac{p_{\mathrm{oI}} - p_{\mathrm{eI}}}{1 - p_{\mathrm{eI}}}. \tag{1.2}$$

Interpretation

Two main criticisms on kappa coefficients were formulated in the literature. First, like correlation coefficients, kappa coefficients vary between −1 and +1 and have no clear interpretation, except for 0 and 1 values. Landis and Koch (1977) proposed qualifying the strength of agreement according to the values taken by the kappa coefficient. This classification is widely used but should be avoided because its construction is totally arbitrary and the value of kappa coefficients depends on the prevalence of the trait studied. It is preferable to consider a confidence interval to appreciate the value of a kappa estimate; often only the lower bound is of interest. Several methods were derived to estimate the sampling variability of kappa-like agreement coefficients. Most of the statistical packages (e.g. SAS, SPSS, STATA, R) report the sample variance given by the delta method. The formula is given in the Appendix for the general case of more than two categories.

Second, several authors pointed out that kappa coefficients are dependent on the prevalence of the trait under study, which indicates a serious limitation when comparing values of kappa coefficients among studies with varying prevalence. More precisely, Thompson and Walter (1988) demonstrated that kappa coefficients can be written as a function of the true prevalence of the trait, as well as the sensitivity and the specificity of each rater classification. This dependence can

lead to surprising results when a high observed proportion of agreement is associated with a low kappa value.

Some alternatives to the classic Cohen's kappa coefficient have been proposed to cope with this problem. For example, the bias-adjusted kappa (BAK) allows adjustment of Cohen's kappa coefficient for rater bias (i.e. differences in the marginal probability distribution of the two raters). The BAK coefficient turns out to be equivalent to the intraclass kappa coefficient $\hat{\kappa}_1$ defined in Eq. 1.2. Furthermore, a prevalence-adjusted-bias-adjusted kappa (PABAK), which is nothing more than a linear transformation of the observed proportion of agreement (PABAK $= 2p_o - 1$), was suggested by Byrt et al. (1993).

Therefore, despite its drawbacks, Cohen's kappa coefficient remains popular to assess agreement in the absence of a gold standard. However, it should be kept in mind that Cohen's kappa coefficient mixes two sources of disagreement among raters: disagreement due to bias among raters (i.e. different probabilities to classify elements in category 1 for the two raters) and disagreement that occurs because the raters evaluate the elements differently (i.e. rank order the elements differently). Rater bias can be studied by comparing values of the kappa coefficient and the intraclass kappa coefficient. Cohen' kappa coefficient is always larger than the intraclass kappa coefficient because it does not penalize for rater bias, equivalence being reached when there is no rater bias ($n_{12} = n_{21}$). Therefore, the larger the difference between the two coefficients, the larger the rater bias. On the other hand, a difference between the intraclass kappa coefficient (BAK) and PABAK indicates that the marginal probability distributions of the raters depart from the uniform distribution ($\bar{p}_1 = \bar{p}_2 = 0.5$).

Example

Consider the cervical ectopy data given in Table 1.1, where two medical raters classify the cervical ectopy size of 85 women. To determine the agreement on each category separately, 2×2 tables were constructed by isolating one category and collapsing all the other categories together (Table 1.3).

When considering the category minimal against all other categories, the observed proportion of agreement is equal to

$$p_o = \frac{13 + 56}{85} = 0.81.$$

This means that the two medical raters classify 81 % of the women in the same category, that is, they agree on 81 % of the women. The proportion of agreement expected by chance is equal to

$$p_e = \frac{27 \times 15 + 58 \times 70}{85^2} = 0.62.$$

Table 1.3 Contingency tables obtained from the classification of the ectopy size of 85 women by two medical raters with direct visual assessment when isolating each category of the four-category scale

Category minimal				Category moderate			
	Rater 2				Rater 2		
Rater 1	Minimal	Other	Total	Rater 1	Moderate	Other	Total
Minimal	13	2	15	Moderate	16	13	29
Other	14	56	70	Other	13	43	56
Total	27	58	85	Total	29	56	85
Category large				**Category excessive**			
	Rater 2				Rater 2		
Rater 1	Large	Other	Total	Rater 1	Excessive	Other	Total
Large	3	10	13	Excessive	11	17	28
Other	15	57	72	Other	0	57	57
Total	18	67	85	Total	11	74	85

Therefore, given the marginal distribution of the two medical raters, if they classify the elements randomly, we expect them to agree on 62 % of the women. This leads to a Cohen's kappa coefficient of

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e} = \frac{0.81 - 0.62}{1 - 0.62} = 0.51.$$

In a same way, the intraclass kappa coefficient is equal to 0.49. The results obtained for the other categories are summarized in Table 1.4.

It can be observed in Table 1.4 that there is a significant positive agreement on all categories, except on category large (the lower bound of the 95 % confidence interval is negative). More generally, it is seen that the agreement on extreme categories (minimal and excessive) is better than the agreement on the middle categories (moderate and large). This is a well-know phenomenon. When the marginal distributions of the two raters are the same (see category moderate in Table 1.3), we have $\hat{\kappa} = \hat{\kappa}_I$, as expected.

Categorical Scale

Cohen's Kappa and Intraclass Kappa Coefficients

Consider now the situation where two raters have to classify N elements on a categorical scale with more than two ($K > 2$) categories (e.g. cervical ectopy size is rated on a four-category scale). By extension, Cohen (1960) defined the *observed proportion of agreement* and the *proportion of agreement expected by chance* by

Table 1.4 Observed proportion of agreement, proportion of agreement expected by chance, kappa coefficient, standard error and 95 % confidence interval (95 % CI) of the Cohen's kappa coefficient and the intraclass kappa coefficient for each 2×2 table given in Table 1.3

	Cohen's kappa				
Category	p_o	p_e	$\hat{\kappa}$	SE($\hat{\kappa}$)	95 % CI
Minimal	0.81	0.62	0.51	0.10	0.31, 0.71
Moderate	0.69	0.55	0.32	0.11	0.11, 0.53
Large	0.71	0.70	0.019	0.11	−0.19, 0.23
Excessive	0.80	0.63	0.47	0.098	0.27, 0.66
	Intraclass kappa				
	p_{oI}	p_{eI}	$\hat{\kappa}_I$	SE($\hat{\kappa}_I$)	95 % CI
Minimal	0.81	0.63	0.49	0.11	0.27, 0.72
Moderate	0.69	0.55	0.32	0.11	0.11, 0.53
Large	0.71	0.70	0.014	0.13	−0.24, 0.27
Excessive	0.80	0.65	0.43	0.11	0.23, 0.64

$$p_o = \sum_{j=1}^{K} \frac{n_{jj}}{N} = \sum_{j=1}^{K} p_{jj} \quad \text{and} \quad p_e = \sum_{j=1}^{K} \frac{n_{j\cdot} n_{\cdot j}}{N^2} = \sum_{j=1}^{K} p_{j\cdot} p_{\cdot j}.$$

This leads to the Cohen's kappa coefficient

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}.$$

In the same way, we have for the intraclass kappa coefficient

$$p_{oI} = \sum_{j=1}^{K} p_{jj}, \quad p_{eI} = \sum_{j=1}^{K} \left(\frac{p_{j\cdot} + p_{\cdot j}}{2} \right)^2 \quad \text{and} \quad \hat{\kappa}_I = \frac{p_{oI} - p_{eI}}{1 - p_{eI}}.$$

Interpretation

It has been proven that Cohen's kappa and the intraclass kappa coefficients computed for a $K \times K$ contingency table are in fact weighted averages of kappa coefficients obtained on 2×2 tables, constructed by isolating a single category $[j]$ from the other categories (see Table 1.3) ($j = 1, \ldots, K$). The overall proportion of observed agreement is in fact the average of the observed proportion of agreement in the 2×2 tables and the same applies for the proportion of agreement expected by chance. More precisely, we have

$$\hat{\kappa} = \frac{\sum_{j=1}^{K} (p_{o[j]} - p_{e[j]})}{\sum_{j=1}^{K} (1 - p_{e[j]})} = \frac{1}{\sum_{j=1}^{K} (1 - p_{e[j]})} \sum_{j=1}^{K} (1 - p_{e[j]}) \hat{\kappa}_{[j]}.$$

Example

In the cervical ectopy example, the proportion of observed agreement and the proportion of agreement expected by chance are respectively equal to

$$p_o = (13 + 16 + 3 + 11)/85 = 0.506$$

and

$$p_e = (27 \times 15 + 29 \times 29 + 18 \times 13 + 11 \times 28)/85^2 = 0.247.$$

Cohen's kappa coefficient is then equal to

$$\hat{\kappa} = \frac{0.506 - 0.247}{1 - 0.247} = 0.34 \quad (95\ \%\ \text{CI}\ 0.21 - 0.48).$$

The average of the observed and expected proportions of agreement in the 2×2 tables (see Table 1.3) are $p_o = (0.81 + 0.69 + 0.71 + 0.80)/4 = 0.506$ and $p_e = (0.62 + 0.55 + 0.70 + 0.63)/4 = 0.247$, as expected.

In the same way, the overall intraclass kappa coefficient is equal to $\hat{\kappa}_I = (0.506 - 0.263)/(1 - 0.263) = 0.33$ (95 % CI 0.19–0.47).

Ordinal Scale

Weighted Kappa Coefficients

Some disagreements between two raters can be considered more important than others. For example, on an ordinal scale, disagreements on two extreme categories are generally considered more important than on neighbouring categories. In the cervical ectopy example, discordance between minimal and excessive has more impact than between minimal and moderate. For this reason, in 1968 Cohen introduced the weighted kappa coefficient. Agreement (w_{jk}) or disagreement (v_{jk}) weights are a priori distributed in the K^2 cells of the $K \times K$ contingency table

summarizing the classification of the two raters, to reflect the seriousness of disagreement according to the distance between the categories. The weighted kappa coefficient is then defined in terms of agreement weights

$$\hat{\kappa}_{\mathrm{w}} = \frac{p_{\mathrm{ow}} - p_{\mathrm{ew}}}{1 - p_{\mathrm{ew}}} \tag{1.3}$$

with

$$p_{\mathrm{ow}} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{jk} \quad \text{and} \quad p_{\mathrm{ew}} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{j\cdot} p_{\cdot k}$$

(usually $0 \leq w_{jk} \leq 1$ and $w_{jj} = 1$), or in terms of disagreement weights

$$\hat{\kappa}_{\mathrm{w}} = 1 - \frac{q_{\mathrm{ow}}}{q_{\mathrm{ew}}} \tag{1.4}$$

with

$$q_{\mathrm{ow}} = \sum_{j=1}^{K} \sum_{k=1}^{K} v_{jk} p_{jk} \quad \text{and} \quad q_{\mathrm{ew}} = \sum_{j=1}^{K} \sum_{k=1}^{K} v_{jk} p_{j\cdot} p_{\cdot k}$$

(usually $0 \leq v_{jk} \leq 1$ and $v_{jj} = 0$).

Although weights can be arbitrarily defined, two agreement weighting schemes defined by Cicchetti and Allison (1971) are commonly used. These are the linear and quadratic weights, given respectively by

$$w_{jk} = 1 - \frac{|j-k|}{K-1} \quad \text{and} \quad w_{jk} = 1 - \left(\frac{|j-k|}{K-1}\right)^2.$$

The disagreement weights $v_{jk} = (j-k)^2$ are also used. Note that Cohen's kappa coefficient is a particular case of the weighted kappa coefficient where $w_{jk} = 1$ when $j = k$ and $w_{jk} = 0$ otherwise.

Interpretation

The use of weighted kappa coefficients was also criticized in the literature, mainly because the weights are generally given a priori and defined arbitrarily. Quadratic weights have received much attention in the literature because of their practical interpretation. For instance, using the disagreement weights $v_{jk} = (j-k)^2$, the weighted kappa coefficient can be interpreted as an intraclass correlation coefficient

in a two-way analysis of variance setting (see Fleiss and Cohen (1973); Schuster (2004)).

By contrast, linear weights possess an intuitive interpretation. The $K \times K$ contingency table can be reduced into a 2×2 classification table by grouping the first k categories in one category and the last $K - k$ categories in a second category $(k = 1, \ldots, K - 1)$. The linearly weighted observed and expected agreements are then merely the mean values of the corresponding proportions of all these 2×2 tables. Therefore, similar to Cohen's kappa coefficient, the linearly weighted kappa coefficient is a weighted average of individual kappa coefficients (see Vanbelle and Albert (2009).

The value of the weighted kappa coefficient can vary considerably for different weighting schemes used and henceforth may lead to different conclusions. Clear guidelines for the selection of weights are not yet available in the literature. However, Warrens (2012) tends to favour the use of the linearly weighted kappa because the quadratically weighted kappa is not always sensitive to changes in the diagonal cells of a contingency table.

Example

Consider again the cervical ectopy size example, where women are classified on a four-category Likert scale by two raters (see Table 1.1). The linear and quadratic agreement weights corresponding to the four-category scale are given in Table 1.5. As an illustration, the linear and quadratic weights for the cell (1,2) are equal to $1 - |1 - 2|/(4 - 1) = 0.67$ and $1 - |1 - 2|^2/(4 - 1)^2 = 0.89$, respectively.

To determine the linearly and quadratically weighted kappa coefficient, we have to determine the weighted observed agreement and weighted expected agreement separately. For each of the $K \times K$ cells, we have to multiply the proportion of elements in the cell (p_{jk}) by the corresponding weight (w_{jk}) and then sum these to obtain p_{ow} (Table 1.6). The weighted expected agreement is obtained similarly. For cell (1,2), we have $w_{12}n_{12}/N = 0.67 \times 2/85 = 0.016$ and $0.89 \times 2/85 = 0.021$, respectively.

The linearly weighted kappa coefficient ($\pm$SE) obtained is 0.52 ± 0.060 (95 % CI 0.40–0.64) with $p_{\text{ow}} = 0.80$ and $p_{\text{ew}} = 0.58$. The quadratically weighted kappa coefficient is 0.67 (95 % CI 0.55–0.78) with $p_{\text{ow}} = 0.91$ and $p_{\text{ew}} = 0.72$. In this example, the quadratically weighted kappa coefficient is greater than the linearly weighted kappa coefficient. However, the reverse could happen in other data sets. No clear relationship between the two coefficients has been established in the literature.