UNIVERSITE DE LIEGE

FACULTE DES SCIENCES
Département de Mathématique

Agreement between raters and groups of raters

Année académique
2008-2009

Dissertation présentée par
Sophie VANBELLE
en vue de l'obtention du grade de
Docteur en Sciences

UNIVERSITE DE LIEGE

FACULTE DES SCIENCES
Département de Mathématique

Agreement between raters and groups of raters

Sophie Vanbelle
June 2009

Aknowledgments

At the term of this doctoral work, I like to express my sincere gratitude to my supervisors, Professors Adelin Albert and Gentiane Haesbroeck. Professor Albert developed my passion for data analysis and simultaneously for scientific research. With his long experience and wisdom, he also promoted this work shortly after my master degree and he encouraged me to proceed patiently and step by step. His comments and remarks greatly improved the text and content of this thesis. Likewise, the encouragements and advice of Professor Haesbroeck contributed to the finalization of this thesis. I also thank the members of the jury for their interest in this work.

I am much indebted to Laetitia Comté, Laurence Seidel, Lixin Zhang and all my colleagues from the Department of Public Health and the Department of Mathematics for the discussions and their daily company. A special word of gratitude goes to Anne-Françoise Donneau for sharing more than the same office with me over the past 6 years.

I cannot escape thanking my family and friends for their support during the ups and downs of my research pathway. Thanks go to my parents, who encouraged me in all my study choices and permitted their realization, but also to my brother, who was always standing by my side. I would also especially thank my companion Jean-Baptiste for his patience and his dedicated attention, my son Antoine, who, as baby, did not wake up at night, leaving me free time for work, and the baby I am carrying now for her discretion at the beginning of the pregnancy. I am also grateful to Julie, Juliette and Roxane, members of "l'ASSOS", for their availability, their uncommon view of life and their taste for monastic beers.

This thesis is dedicated to my son Antoine.

Be not afraid of growing slowly,
be afraid only of standing still.
(Chinese proverb)

Summary

Agreement between raters on a categorical scale is not only a subject of scientific research but also a problem frequently encountered in practice. Whenever a new scale is developed to assess individuals or items in a certain context, inter-rater agreement is a prerequisite for the scale to be actually implemented in routine use. Cohen's kappa coefficient is a landmark in the developments of rater agreement theory. This coefficient, which operated a radical change in previously proposed indexes, opened a new field of research in the domain.

In the first part of this work, after a brief review of agreement on a quantitative scale, the kappa-like family of agreement indexes is described in various instances: two raters, several raters, an isolated rater and a group of raters and two groups of raters. To quantify the agreement between two individual raters, Cohen's kappa coefficient (Cohen, 1960) and the intraclass kappa coefficient (Kraemer, 1979) are widely used for binary and nominal scales, while the weighted kappa coefficient (Cohen, 1968) is recommended for ordinal scales. An interpretation of the quadratic (Schuster, 2004) and the linear (Vanbelle and Albert, 2009c) weighting schemes is given. Cohen's kappa (Fleiss, 1971) and intraclass kappa (Landis and Koch, 1977c) coefficients were extended to the case where agreement is searched between several raters. Next, the kappa-like family of agreement coefficients is extended to the case of an isolated rater and a group of raters (Vanbelle and Albert, 2009a) and to the case of two groups of raters (Vanbelle and Albert, 2009b). These agreement coefficients are derived on a population-based model and reduce to the well-known Cohen's kappa coefficient in the case of two single raters. The proposed agreement indexes are also compared to existing methods, the consensus method and Schouten's agreement index (Schouten, 1982). The superiority of the

new approach over the latter is shown.

In the second part of the work, methods for hypothesis testing and data modeling are discussed. Firstly, the method proposed by Fleiss (1981) for comparing several independent agreement indexes is presented. Then, a bootstrap method initially developed by McKenzie et al. (1996) to compare two dependent agreement indexes, is extended to several dependent agreement indexes (Vanbelle and Albert, 2008). All these methods equally apply to the kappa coefficients introduced in the first part of the work. Next, regression methods for testing the effect of continuous and categorical covariates on the agreement between two or several raters are reviewed. This includes the weighted least-squares method allowing only for categorical covariates (Barnhart and Williamson, 2002) and a regression method based on two sets of generalized estimating equations. The latter method was developed for the intraclass kappa coefficient (Klar et al., 2000), Cohen's kappa coefficient (Williamson et al., 2000) and the weighted kappa coefficient (Gonin et al., 2000). Finally, a heuristic method, restricted to the case of independent observations, is presented (Lipsitz et al., 2001, 2003) which turns out to be equivalent to the generalized estimating equations approach. These regression methods are compared to the bootstrap method extended by Vanbelle and Albert (2008) but they were not generalized to agreement between a single rater and a group of raters nor between two groups of raters.

Résumé

Sujet d'intenses recherches scientifiques, l'accord entre observateurs sur une échelle catégorisée est aussi un problème fréquemment rencontré en pratique. Lorsqu'une nouvelle échelle de mesure est développée pour évaluer des sujets ou des objets, l'étude de l'accord inter-observateurs est un prérequis indispensable pour son utilisation en routine. Le coefficient kappa de Cohen constitue un tournant dans les développements de la théorie sur l'accord entre observateurs. Ce coefficient, radicalement différent de ceux proposés auparavant, a ouvert de nouvelles voies de recherche dans le domaine.

Dans la première partie de ce travail, après une brève revue des mesures d'accord sur une échelle quantitative, la famille des coefficients kappa est décrite dans différentes situations: deux observateurs, plusieurs observateurs, un observateur isolé et un groupe d'observateurs, et enfin deux groupes d'observateurs. Pour quantifier l'accord entre deux observateurs, le coefficient kappa de Cohen (Cohen, 1960) et le coefficient kappa intraclasse (Kraemer, 1979) sont largement utilisés pour les échelles binaires et nominales. Par contre, le coefficient kappa pondéré (Cohen, 1968) est recommandé pour les échelles ordinales. Schuster (2004) a donné une interprétation des poids quadratiques tandis que Vanbelle and Albert (2009c) se sont interessés aux poids linéaires. Les coefficients d'accord correspondant au coefficient kappa de Cohen (Fleiss, 1971) et au coefficient kappa intraclasse (Landis and Koch, 1977c) sont aussi donnés dans le cas de plusieurs observateurs. La famille des coefficients kappa est ensuite étendue au cas d'un observateur isolé et d'un groupe d'observateurs (Vanbelle and Albert, 2009a) et au cas de deux groupes d'observateurs (Vanbelle and Albert, 2009b). Les coefficients d'accord sont élaborés à partir d'un modèle de population et se réduisent au coefficient

kappa de Cohen dans le cas de deux observateurs isolés. Les coefficients d'accord proposés sont aussi comparés aux méthodes existantes, la méthode du consensus et le coefficient d'accord de Schouten (Schouten, 1982). La supériorité de la nouvelle approche sur ces dernières est démontrée.

Des méthodes qui permettent de tester des hypothèses et modéliser des coefficients d'accord sont abordées dans la seconde partie du travail. Une méthode permettant la comparaison de plusieurs coefficients d'accord indépendants (Fleiss, 1981) est d'abord présentée. Puis, une méthode basée sur le bootstrap, initialement développée par McKenzie et al. (1996) pour comparer deux coefficients d'accord dépendants, est étendue au cas de plusieurs coefficients dépendants par Vanbelle and Albert (2008). Pour finir, des méthodes de régression permettant de tester l'effet de covariables continues et catégorisées sur l'accord entre deux observateurs sont exposées. Ceci comprend la méthode des moindres carrés pondérés (Barnhart and Williamson, 2002), admettant seulement des covariables catégorisées, et une méthode de régression basée sur deux équations d'estimation généralisées. Cette dernière méthode a été développée dans le cas du coefficient kappa intraclasse (Klar et al., 2000), du coefficient kappa de Cohen (Williamson et al., 2000) et du coefficient kappa pondéré (Gonin et al., 2000). Enfin, une méthode heuristique, limitée au cas d'observations indépendantes, est présentée (Lipsitz et al., 2001, 2003). Elle est équivalente à l'approche par les équations d'estimation généralisées. Ces méthodes de régression sont comparées à l'approche par le bootstrap (Vanbelle and Albert, 2008) mais elles n'ont pas encore été généralisées au cas d'un observateur isolé et d'un groupe d'observateurs ni au cas de deux groupes d'observateurs.

Samenvatting

Het bepalen van overeenstemming tussen beoordelaars voor categorische gegevens is niet alleen een kwestie van wetenschappelijk onderzoek, maar ook een probleem dat men veelvuldig in de praktijk tegenkomt. Telkens wanneer een nieuwe schaal wordt ontwikkeld om individuele personen of zaken te evalueren in een bepaalde context, is interbeoordelaarsovereenstemming een noodzakelijke voorwaarde vooraleer de schaal in de praktijk kan worden toegepast. Cohen's kappa coëfficiënt is een mijlpaal in de ontwikkeling van de theorie van interbeoordelaarsovereenstemming. Deze coëfficiënt, die een radicale verandering met de voorgaande indices inhield, opende een nieuw onderzoeksspoor in het domein.

In het eerste deel van dit werk wordt, na een kort overzicht van overeenstemming voor kwantitatieve gegevens, de kappa-achtige familie van overeenstemmingsindices beschreven in verschillende gevallen: twee beoordelaars, verschillende beoordelaars, één geïsoleerde beoordelaar en een groep van beoordelaars, en twee groepen van beoordelaars. Om de overeenstemming tussen twee individuele beoordelaars te kwantificeren worden Cohen's kappa coëfficiënt (Cohen, 1960) en de intraklasse kappa coëfficiënt (Kraemer, 1979) veelvuldig gebruikt voor binaire en nominale gegevens, terwijl de gewogen Kappa coëfficiënt (Cohen, 1968) aangewezen is voor ordinale gegevens. Een interpretatie van de kwadratische (Schuster, 2004) en lineaire (Vanbelle and Albert, 2009c) weegschema's wordt gegeven. Overeenstemmingsindices die overeenkomen met Cohen's Kappa (Fleiss, 1971) en intraklassekappa (Landis and Koch, 1977c) coëfficiënten kunnen worden gebruikt om de overeenstemming tussen verschillende beoordelaars te beschrijven. Daarna wordt de familie van kappa-achtige overeenstemmingscoëfficiënten uitgebreid tot het geval van één geïsoleerde beoordelaar en een groep van beoordelaars (Vanbelle and

Albert, 2009a) en tot het geval van twee groepen van beoordelaars (Vanbelle and Albert, 2009b). Deze overeenstemmingscoëfficiënten zijn afgeleid van een populatie-gebaseerd model en kunnen worden herleid tot de welbekende Cohen's coëfficiënt in het geval van twee individuele beoordelaars. De voorgestelde overeenstemmingsindices worden ook vergeleken met bestaande methodes, de consensusmethode en Schoutens overeenstemmingsindex (Schouten, 1982). De superioriteit van de nieuwe benadering over de laatstgenoemde wordt aangetoond.

In het tweede deel van het werk worden hypothesetesten en gegevensmodellering besproken. Vooreerst wordt de methode voorgesteld door Fleiss (1981) om verschillende onafhankelijke overeenstemmingsindices te vergelijken, voorgesteld. Daarna wordt een bootstrapmethode, oorspronkelijk ontwikkeld door McKenzie et al. (1996) om twee onafhankelijke overeenstemmingsindices te vergelijken, uitgebreid tot verschillende afhankelijke overeenstemmingsindices (Vanbelle and Albert, 2008). Al deze methoden kunnen ook worden toegepast op de overeenstemmingsindices die in het eerste deel van het werk zijn beschreven. Ten slotte wordt een overzicht gegeven van regressiemethodes om het effect van continue en categorische covariabelen op de overeenstemming tussen twee of meer beoordelaars te testen. Dit omvat de gewogen kleinste kwadraten methode, die alleen werkt met categorische covariabelen (Barnhart and Williamson, 2002) en een regressiemethode gebaseerd op twee sets van gegeneraliseerde schattingsvergelijkingen. De laatste methode was ontwikkeld voor de intraklasse kappa coëfficiënt (Klar et al., 2000), Cohen's kappa coëfficiënt (Williamson et al., 2000) en de gewogen kappa coëfficiënt (Gonin et al., 2000). Ten slotte wordt een heuristische methode voorgesteld die alleen van toepassing is op het geval van onafhankelijk waarnemingen (Lipsitz et al., 2001, 2003). Ze blijkt equivalent te zijn met de benadering van de gegeneraliseerde schattingsvergelijkingen. Deze regressiemethoden worden vergeleken met de bootstrapmethode uitgebreid door Vanbelle and Albert (2008) maar werden niet veralgemeend tot de overeenstemming tussen een enkele beoordelaar en een groep van beoordelaars, en ook niet tussen twee groepen van beoordelaars.

Contents

Glossary

ANOVA	Analysis of Variance
BL	Borderline
BMS	Between items Mean Squares
BSS	Between items Sum of Squares
CAP	College of American Pathologists
CCC	Concordance Correlation Coefficient
CI	Confidence Interval
DVT	Deep venous thrombosis
EMIT	Enzyme-mediated immunoassay technique
EMS	Error Mean Square
ESS	Error Sum of Squares
FIA	Fluoro-immunoassay
FTA-ABS	Fluorescent treponemal antibody absorption test
GEE	Generalized estimating equations
GEE1	First order generalized estimating equations
GEE2	Second order generalized estimating equations
GLM	Generalized linear models
GLMM	Generalized linear mixed models
GSK	Grizzle, Starmer and Koch methodology
HIV	Human immunodeficiency virus

ICC	Intraclass Correlation Coefficient
ICC_A	Agreement Intraclass Correlation Coefficient
ICC_C	Consistency Intraclass Correlation Coefficient
JMS	Between raters Mean Squares
JSS	Between raters Sum of Squares
MDCT	Multidetector-row computed tomography
MLE	Maximum likelihood estimator
MS	Mean squares
NR	Non reactive
RE	Reactive
SCT	Script concordance test
SD	Standard deviation
SE	Standard error
TCC	Tetrachoric correlation coefficient
TSS	Total Sum of Squares
US	Ultrasound
WLS	Weighted Least Squares
WMS	Within items Mean Squares
WSS	Within items Sum of Squares

General introduction

Reliable and accurate measurements serve as the basis for evaluation in social, medical, behavioral and biological sciences (Barnhart et al., 2007). As new concepts, theories and technologies continue to develop, new scales, methods, tests, assays and instruments become available for the evaluation. Since errors are inherent to every measurement procedure, one must ensure that the measurement is accurate before it is used in practice. In simple intuitive terms, a reliable and accurate measurement may simply mean that the new measurement is the same as the truth or agree with the truth. However, requiring the new measurement to be identical to the truth is often impracticable because we are willing to accept a measurement up to some tolerable error or because the truth is simply not available to us. To deal with these issues, a number of theoretical and methodological approaches has been proposed over the years in different disciplines. A vast literature is covering aspects related to the concordance between quantitative scales, in particular in methods comparisons. A classical example is laboratory medicine, where any new analytical technique or instrument needs to be compared to the routine one before it is actually implemented in practice. We shall briefly review this topic although it is not the major focus of the present work.

This work rather focuses on agreement between raters on a categorical scale. The most elementary situation concerns agreement assessment between two raters on a binary scale. For example, we may be interested in the agreement between two radiologists (say, a junior one and a senior one) in visualizing patient x-rays and classifying them as normal or abnormal, or in the agreement between two scientific experts judging separately a series of grant applications as accepted or rejected. Clearly in both examples, we would hope that raters agree to a large extent. Unfortunately, agreement can occur by chance alone. Thus, in the examples above,

if the two radiologists or the two experts toss a coin for each item to be classified rather than doing their job, there will be a non negligible number of cases where the coin toss will give the same outcome.

Cohen (1960) was the first to recognize this fact which led him to introduce the celebrated kappa coefficient, also known as Cohen's kappa coefficient. The latter has been widely used ever since. The extension of Cohen's kappa coefficient between two raters for categorical scales was straightforward and followed the same principle as for the dichotomous scale. Categorical scales are widely used in psychometry, as for instance the well-known Likert scale. The agreement between several raters appeared as a natural extension of the two raters problem but raised a number of new issues that had to be tackled. Approaches similar to those available for quantitative scales were developed, leading to the definition of so-called intraclass coefficients (Fleiss, 1971; Davies and Fleiss, 1982). Landis and Koch (1975a,b) made a comprehensive review of the various agreement indexes between two or more raters used for categorical scales.

There are situations where agreement is searched between an isolated rater and a group of raters, or between two groups of raters. For instance, in medical education and even more generally, it is common to assess the knowledge level of the students by challenging them against a group of experts. The Script Concordance Test (SCT) proposed by Charlin et al. (2002) is one way to do this assessment. Although our personal interest for agreement coefficients arose with our master thesis (Vanbelle, 2002), the SCT application really motivated our research work because existing solutions were not satisfactory.

The present work is divided in seven chapters. Chapter 1 gives a brief overview of agreement measures for quantitative scales. After describing two graphical methods (e.g., Bland and Altman plot), we present the concordance correlation coefficient (CCC) introduced by Lin (1989) which quantifies the agreement between two raters for quantitative data. Then we move to the intraclass correlation coefficients (ICC), allowing to assess quantification of agreement between several raters. The description of the various agreement coefficients is limited to the most simplest cases and particular emphasis is placed on aspects that were used later on for qualitative scales.

In Chapter 2, kappa-like agreement indexes are reviewed to quantify the agreement between two raters on a categorical scale. This includes Cohen's kappa coefficient (Cohen, 1960), the intraclass kappa coefficient (Kraemer, 1979) and the weighted kappa coefficient (Cohen, 1968). Interpretation of the weights is provided for the two most used weighting schemes: the linear (Vanbelle and Albert, 2009c) and the

quadratic (Schuster, 2004) weighting schemes. The asymptotic sampling variance of the agreement indexes is also considered.

Chapter 3 generalizes the agreement indexes introduced in Chapter 2 to the case of several raters. These agreement indexes are mostly based on linear ANOVA models and mimic the intraclass correlation coefficients introduced for quantitative scales (Landis and Koch, 1977c) or are based on pairwise agreement (Davies and Fleiss, 1982). All agreement indexes are given for both binary and multinomial scales.

Novel extensions of agreement coefficients described in Chapter 2 are dealt with in chapters 4 and 5. They constitute the salient core of this work. The agreement problem between an isolated rater and a group of raters is discussed in depth in Chapter 4, whereas the agreement between two groups of raters is the subject topic of Chapter 5. New agreement indexes are proposed (Vanbelle and Albert, 2009a,b) and compared to the consensus method, known to be unsatisfactory, and to the more general method developed by Schouten (1982).

Hypothesis testing methods on kappa coefficients are described in Chapter 6. A distinction is made between tests on a single kappa coefficient and tests on several kappa coefficients. When comparing several kappa coefficients, a further distinction is made between independent (unpaired case) coefficients (Fleiss, 1981) and dependent (paired case) coefficients (McKenzie et al., 1996; Vanbelle and Albert, 2008).

Finally, Chapter 7 is devoted to recent advances on kappa coefficients in the context of generalized linear mixed models (GLMM). Theses approaches permit the modeling of agreement indexes according to covariates. This includes the weighted least squares (Barnhart and Williamson, 2002) approach and the generalized estimating equations (Klar et al., 2000). Their performance are compared to ours on a couple of examples (Vanbelle and Albert, 2008).

In summary, the present work intends to provide a comprehensive overview of the problem of rater agreement, which hopefully could serve as a reference text for any scientist interested in the domain. We have incorporated our personal original research findings in a more general framework in order to present a global and coherent view of past developments and recent advances in the problem of rater agreement.

CHAPTER 1

Agreement on a quantitative scale

1.1 Introduction

From a statistical standpoint, the problem of agreement on a quantitative scale has been a subject of interest before that of agreement on a qualitative scale and some of the methods developed for quantitative measurements were adapted to the case of categorical observations. When measuring a quantity with a new instrument, two questions typically arise: (1) is the new instrument calibrated against the established method, and (2) are the measurements made with the new instrument reproducible? The established method is often regarded as a 'gold standard' or reference method measuring the "true" value of the quantity to be determined. However, when comparing two methods, it is frequent that none of them can be viewed as giving a true value. Then, an assessment of the degree of agreement between the two methods is required to evaluate the comparability of the measurements. In practice, the measurements obtained with the two methods can be plotted on a 2-dimensional graph, perfect agreement occurring when all measurements fall on the 45° line. Another approach (Bland and Altman, 1986) is to display the difference of the two measurements against their mean in which case perfect agreement would correspond to all points laying on the abscissa. These methods are basically visual. Lin (1989) therefore introduced the concordance correlation coefficient (CCC) measuring the correlation between duplicate measurements falling around the 45° line through the origin. The CCC was generalized to more than two measurement methods in various situations. The interested reader can refer to Barnhart et al. (2007) for a complete overview. As an

alternative approach, a methodology based on the analysis of variance (ANOVA) was developed by Fisher (1958), leading to the intraclass correlation coefficient (ICC) which is a reliability criterion giving the proportion of variance attributable to differences between methods. The ICC was developed to deal with several measurement methods and has emerged as a universal and widely accepted reliability index (Shoukri, 2004). Several versions of the ICC were derived depending on the study scheme (Bartko, 1966; Shrout and Fleiss, 1979). In this chapter, we shall restrict our overview of agreement indexes on a quantitative scale to those which have been extended to qualitative scales.

1.2 Agreement between two raters

1.2.1 Visual assessment of agreement

Let Y denote a quantity associated with each element (item) of an infinite population $\mathcal{I}$. For simplicity, let $Y \geq 0$. Further, let Y_1 and Y_2 denote the corresponding quantities as measured by two distinct raters. In theory, perfect agreement between the raters occurs when $Y_1 = Y_2$. Thus, given a sample of items, the agreement between the two raters is best seen by plotting the paired observations with respect to the 45° line ($Y_2 = Y_1$). If the two raters are in perfect agreement, all observations will fall on the 45° line. By contrast, disagreement between the two raters can take different forms: (i) a constant bias ($Y_2 = a + Y_1$, $a \in \mathbb{R}$), (ii) a proportional bias ($Y_2 = bY_1$, $b \in \mathbb{R}$), or (iii) both types of biases ($Y_2 = a + bY_1$, $a, b \in \mathbb{R}$). This led Bland and Altman (1986) to suggest plotting the difference of the measurements ($Y_2 - Y_1$) against their mean $(Y_1 + Y_2)/2$, the so-called "Bland and Altman plot". In case of constant bias, $Y_2 - Y_1 = a$ ($a \neq 0$), the observations will tend to lie around a horizontal line; in case of proportional bias, $Y_2 - Y_1 = (b-1)Y_1$ ($b \neq 0$), the points will be scattered around an increasing ($b > 1$) or decreasing ($b < 1$) line passing through the origin; when both biases are present, the increasing or decreasing line will not pass through the origin. Bland and Altman plot can also reveal whether agreement is item related. For instance, two raters may agree closely in estimating the size of small items, but disagree about larger items. The two methods described are essentially graphical, although by regression analysis it is possible to estimate the constant and proportional bias factors.

Example. Shrout and Fleiss (1979) considered the following hypothetical example, where 4 raters measured 6 items on a 10-point scale (see Table 1.1). Consider only the measurements of raters 1 and 4.

As seen on the 45° line plot and on Bland and Altman plot (see Figure 1.1), perfect agreement occurs only for one item (item 3). Moreover, it appears that rater 1

Table 1.1. Example of Shrout and Fleiss (1979)

	Rater					
Item	1	2	3	4	$Y_1 - Y_4$	$(Y_1 + Y_4)/2$
1	9	2	5	8	1	8.5
2	6	1	3	2	4	4
3	8	4	6	8	0	8
4	7	1	2	6	1	6.5
5	10	5	6	9	1	9.5
6	6	2	4	7	-1	6.5

gives in general higher values than rater 4. The mean difference ($\pm SD$) is 1.0±1.67 (95% CI: [-2.28;4.28]).

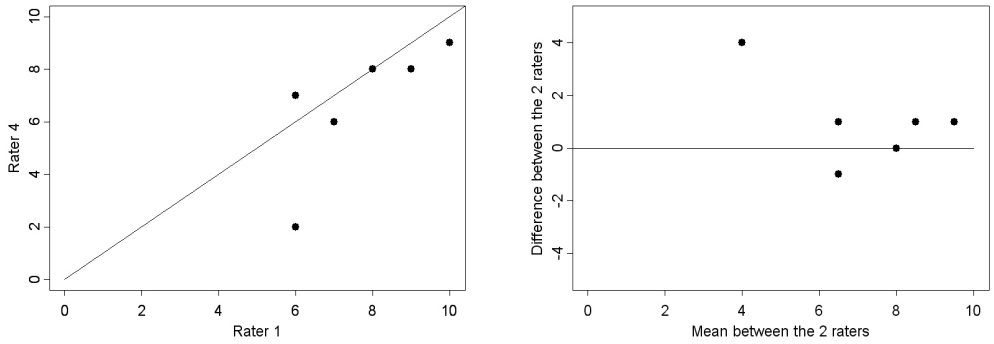

Figure 1.1. 45° line plot (left) and Bland and Altman plot (right) for the measurements of raters 1 and 4 of 6 items on a 10-point scale

1.2.2 Concordance correlation coefficient

There is the need to derive an index reflecting the agreement between the two raters. The recourse to Pearson's correlation coefficient, paired t-test, least-squares analysis of slope and intercept or to the coefficient of variation is always failing in some cases, as shown by Lin (1989). This led Lin (1989) to develop the concordance correlation coefficient (CCC), a reproducibility index measuring the correlation between two readings that fall on the 45° line through the origin.

Definition. Suppose that the joint distribution of Y_1 and Y_2 is bivariate Normal with mean $(\mu_1, \mu_2)'$ and variance-covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}.$$

The degree of concordance between Y_1 and Y_2 can be characterized by the expected value of the squared difference

$$\begin{aligned} E(Y_1 - Y_2)^2 &= (\mu_1 - \mu_2)^2 + (\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}) \\ &= (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2(1-\rho)\sigma_1\sigma_2 \end{aligned} \tag{1.1}$$

where $\rho = corr(Y_1, Y_2) = \sigma_{12}/\sigma_1\sigma_2$. Lin (1989) proposed to apply a transformation to scale the agreement index between -1 and 1, leading to the concordance correlation coefficient

$$CCC = 1 - \frac{E(Y_1 - Y_2)^2}{(\mu_1 - \mu_2)^2 + \sigma_1^2 + \sigma_2^2} = \frac{2\rho\sigma_1\sigma_2}{(\mu_1 - \mu_2)^2 + \sigma_1^2 + \sigma_2^2} = \rho C_b \tag{1.2}$$

where $C_b = [(\nu + 1/\nu + u^2)/2)]^{-1}$ with $\nu = \sigma_1/\sigma_2$ representing the scale shift and $u = (\mu_1 - \mu_2)/\sqrt{\sigma_1\sigma_2}$ the location shift relative to the scale. Lin (1989) noted that C_b ($0 < C_b \leq 1$) is a bias correction factor measuring how far the best-fit deviates from the 45° line (measure of accuracy). No deviation occurs when $C_b = 1$. The Pearson's correlation coefficient ρ measures how far observations deviate from the best-fit line (measure of precision).

The concordance correlation coefficient possesses the following properties:

1. $-1 \leq -|\rho| \leq CCC \leq |\rho| \leq 1$;
2. $CCC = 0$ if and only if $\rho = 0$;
3. $CCC = \rho$ if and only if $\sigma_1 = \sigma_2$ and $\mu_1 = \mu_2$;
4. $CCC = \pm 1$ if and only if each pair of measurements is in perfect agreement or perfect reverse agreement.

Estimation of the parameters. For a sample of N independent pairs $(y_{i,1}, y_{i,2})$, if $\overline{y}_{.,r}$ denotes the estimated mean and s_r^2 the sample variance of the measurements made by rater r ($r = 1, 2$),

$$\overline{y}_{.,r} = \frac{1}{N}\sum_{i=1}^{N} y_{i,r} \quad \text{and} \quad s_r^2 = \frac{1}{N}\sum_{i=1}^{N}(y_{i,r} - \overline{y}_{.,r})^2, \tag{1.3}$$

and if $\hat{\rho}$ is the sample Pearson's correlation coefficient

$$\hat{\rho} = \frac{\sum_{i=1}^{N}(y_{i,1} - \overline{y}_{.,1})(y_{i,2} - \overline{y}_{.,2})/N}{s_1 s_2}, \tag{1.4}$$

the CCC is estimated by

$$\widehat{CCC} = \frac{2\hat{\rho}s_1 s_2}{(\overline{y}_{.,1} - \overline{y}_{.,2})^2 + s_1^2 + s_2^2}. \tag{1.5}$$

Sampling variability. When sampling from a bivariate normal distribution, Lin (1989) showed that $\widehat{CCC}$ has an asymptotic Normal distribution with mean CCC and variance

$$\begin{aligned} var(\widehat{CCC}) &= \frac{1}{N-2}[(1-\rho^2)CCC^2(1-CCC^2)/\rho^2 + 4CCC^3(1-CCC)u^2/\rho \\ &- 2CCC^4u^4/\rho^2]. \end{aligned} \tag{1.6}$$

Example. Consider again the measurements of raters 1 and 4 in the hypothetical example of Shrout and Fleiss (1979) (see Table 1.2).

Table 1.2. Measurements of raters 1 and 4 in the example of Shrout and Fleiss (1979)

	Rater	
Item	1	4
1	9	8
2	6	2
3	8	8
4	7	6
5	10	9
6	6	7
$\overline{y}_{.,r}$	7.7	6.7
s_r	1.63	2.50

We have $\hat{\rho} = 0.75$ and thus,

$$\widehat{CCC} = \frac{2 \times 0.75 \times 1.63 \times 2.50}{(7.7-6.7)^2 + 1.63^2 + 2.50^2} = 0.62$$

with $\hat{C}_b = 0.82$, $\hat{\nu} = 0.65$ and $\hat{u} = 0.49$. We have

$$\begin{aligned} var(\widehat{CCC}) &= \frac{1}{6-2}\left\{\frac{(1-0.75^2)0.62^2(1-0.75^2)}{0.75^2} + \frac{4 \times 0.62^3(1-0.62)0.49^2}{0.75}\right. \\ &- \left.\frac{2 \times 0.62^4 0.49^4}{0.75^2}\right\} = 0.067. \end{aligned}$$

The lower bound of the one-sided 95% confidence interval for the CCC is equal to $0.62 - 1.64\sqrt{(}0.067) = 0.20$. Thus, with 95% confidence $CCC \geq 0.20$.

1.3 Agreement between several raters

The intraclass correlation coefficient (ICC) introduced by Fisher (1958) is universally used as a reliability index. There are several versions of the ICC depending of

the study design (Bartko, 1966), all based on the analysis of variance and the estimation of several variance components. The use of ICC should be restricted by the underlying model which most adequately describes the experiment situation and the conceptual interest of the study. The guidelines for choosing an appropriate form of the ICC are given in Shrout and Fleiss (1979). They suggested determining three important issues to choose an appropriate model: (1) Is one-way or two-way analysis of variance appropriate for the analysis of the reliability study? (2) Are differences between the rater's mean readings relevant to the reliability study? (3) Is the unit of the analysis an individual rating or the mean of several ratings?

Typically, in inter-rater reliability studies, each of a random sample of N items from a population of items $\mathcal{I}$ is rated independently by R raters belonging to a population of raters $\mathcal{R}$. Three different study designs are considered:

Model 1. Each item is rated by a different set of R raters, randomly selected from a larger population of raters. This leads to a one-way random effects ANOVA model.

Model 2. Each item is rated by the same random sample of R raters selected from a larger population. This leads to a two-way random effects ANOVA model with interaction.

Model 3. Each item is rated by each of the same R raters, who are the only raters of interest. This leads to a two-way mixed effects ANOVA model with interaction.

Each kind of study thus requires a separately specified mathematical model to describe its results. The model specifies the decomposition of a measurement made by rater r ($r = 1, \cdots, R$) on item i ($i = 1, \cdots, N$) in terms of various effects. Among the possible effects are the overall effect and the effects for rater r, item i, the interaction between raters and items and for a random error component. Depending on the study design, different effects are estimable, different assumptions must be made about the estimable effects and thus different structures of the ANOVA model are obtained. McGraw and Wong (1996) also distinguished between intraclass correlation coefficients measuring consistency (denoted by ICC_C) and absolute agreement (denoted by ICC_A). The first type of coefficients excludes the variance term relative to the raters from the denominator while the second type does not. To illustrate the distinction between the two kinds of ICC, consider two measurements on a series of items with the second measurement always 2 units higher than the first one. These paired measurements are in perfect agreement using the consistency definition (ICC_C) but not using the absolute agreement definition (ICC_A). The absolute agreement is sensitive to scale shifts while the

consistency agreement is not.

1.3.1 One-way random effects ANOVA model

Suppose that each of a random sample of items $(1, \cdots, N)$ is rated by a different set of raters $(1, \cdots, R_i,\ i = 1, \cdots, N)$. The case of a constant number of ratings $(R_i = R,\ i = 1, \cdots, N)$ for each item is first considered. When each item is rated by a different set of R raters, randomly selected from a larger population of raters (Model 1), the effect due to raters, to the interaction between raters and items and to random error can not be estimated separately. Therefore, only the absolute agreement is measurable. If $Y_{i,r}$ denotes the measurement of rater r $(r = 1, \cdots, R)$ on item i $(i = 1, \cdots, N)$, the following linear model is assumed,

$$Y_{i,r} = \mu + B_i + W_{i,r},\ (i = 1, \cdots, N; r = 1, \cdots, R) \tag{1.7}$$

where μ is the overall population mean of the measurements, B_i is the deviation of item i from μ and $W_{i,r}$ is a residual component equal to the sum of the non separable effects of the raters, the interaction between the raters and the items and the error term.

It is assumed that the component $B_i \sim N(0, \sigma_B^2)$ $(i = 1, \cdots, N)$, the component $W_{i,r} \sim N(0, \sigma_W^2)$ $(i = 1, \cdots, N; r = 1, \cdots, R)$ and that the B_i are independent of $W_{i,r}$. The expected mean squares related to the one-way random effects ANOVA model are given in Table 1.3.

Table 1.3. One-way random effects ANOVA model (Model 1)

Variability	Sum of squares	Degrees of freedom	Mean squares	E(MS)
Between items	BSS	$N-1$	BMS	$R\sigma_B^2 + \sigma_W^2$
Within items	WSS	$N(R-1)$	WMS	σ_W^2
Total	TSS	$NR-1$		

One can see in Table 1.3 that WMS is an unbiased estimate of σ_W^2 and $(BMS - WMS)/R$ is an unbiased estimate of σ_B^2. The intraclass correlation coefficient (ICC_{A1}) is defined by

$$ICC_{A1} = \frac{cov(Y_{i,r}, Y_{i,s})}{\sqrt{var(Y_{i,r})var(Y_{i,s})}} = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2}. \tag{1.8}$$

Estimation of the parameters. The intraclass correlation coefficient ICC_{A1} is estimated by

$$\widehat{ICC}_{A1} = \frac{BMS - WMS}{BMS + (R-1)WMS}. \tag{1.9}$$

This estimate is consistent but biased (Olkin and Pratt, 1958) since the expectation of a ratio is not equal to the ratio of the expectations. Let $y_{i,r}$ denote the observed value of the random variable $Y_{i,r}$ $(i = 1, \cdots, N; r = 1, \cdots, R)$, $\overline{y}_{i,.}$ the mean value over the raters and $\overline{y}_{.,.}$ the overall mean, i.e.,

$$\overline{y}_{i,.} = \frac{1}{R}\sum_{r=1}^{R} y_{i,r}, \text{ and } \overline{y}_{.,.} = \frac{1}{NR}\sum_{i=1}^{N}\sum_{r=1}^{R} y_{i,r}.$$

We have

$$\begin{aligned} BSS &= R\sum_{i=1}^{N}(\overline{y}_{i,.} - \overline{y}_{.,.})^2, \\ WSS &= \sum_{i=1}^{N}\sum_{r=1}^{R}(y_{i,r} - \overline{y}_{i,.})^2, \\ TSS &= \sum_{i=1}^{N}\sum_{r=1}^{R}(y_{i,r} - \overline{y}_{.,.})^2. \end{aligned} \tag{1.10}$$

Confidence interval. Note that

$$\widehat{ICC}_{A1} = \frac{BMS - WMS}{BMS + (R-1)WMS} = \frac{F_0 - 1}{F_0 + (R-1)} \tag{1.11}$$

where $F_0 = BMS/WMS$ is the usual variance ratio distributed as a Snedecor F with $N-1$ and $N(R-1)$ degrees of freedom since B_i and $W_{i,r}$ are normally distributed. If $Q_F(1-\alpha;\nu_1,\nu_2)$ denotes the $(1-\alpha)$-percentile of the F distribution with ν_1 and ν_2 degrees of freedom, then

$$\frac{F_L - 1}{F_L + (R-1)} < ICC_{A1} < \frac{F_U - 1}{F_U + (R-1)} \tag{1.12}$$

is a $(1-\alpha)100\%$ confidence interval for the intraclass correlation coefficient, ICC_{A1}, with

$$\begin{aligned} F_L &= F_0/Q_F(1-\frac{\alpha}{2}; N-1, N(R-1)) \\ \text{and } F_U &= F_0 Q_F(1-\frac{\alpha}{2}; N(R-1), N-1). \end{aligned}$$

Note that in practice, only the lower bound of the confidence interval is usually of interest.

Unequal number of ratings per item. Suppose now that each of a random sample of items $(1, \cdots, N)$ is rated by a different set of raters $(1, \cdots, R_i)$, where

R_i is not the same for all items. In that case, we have

$$\begin{aligned} BSS &= \sum_{i=1}^{N}\sum_{r=1}^{R_i}(\overline{y}_{i,\cdot} - \overline{y}_{\cdot,\cdot})^2, \\ WSS &= \sum_{i=1}^{N}\sum_{r=1}^{R_i}(y_{i,r} - \overline{y}_{i,\cdot})^2, \\ TSS &= \sum_{i=1}^{N}\sum_{r=1}^{R_i}(y_{i,r} - \overline{y}_{\cdot,\cdot})^2. \end{aligned} \tag{1.13}$$

with

$$E(BSS) = R_0\sigma_B^2 + \sigma_W^2 \tag{1.14}$$

where

$$R_0 = \frac{(NR)^2 - \sum_{i=1}^{N} R_i^2}{(N-1)NR}. \tag{1.15}$$

The estimation of the intraclass correlation coefficient ICC_{A1} is then

$$\widehat{ICC}_{A1} = \frac{BMS - WMS}{BMS + (R_0 - 1)WMS}. \tag{1.16}$$

The reader interested by the proofs in case of unequal number of ratings per item may refer to Vanbelle (2002).

Example. Consider again the 4 raters measuring 6 items on the 10-point scale (see Table 1.1) and suppose that the measurements on each item are made by a different set of 4 raters. This leads to the following ANOVA table (Table 1.4). The estimated intraclass correlation coefficient is $\widehat{ICC}_{A1} = 0.17$. We have $F_0 = 1.79$ leading to $F_L = 1.79/2.77 = 0.65$. The one-sided 95% lower bound is equal to -0.10. Thus, there is no evidence for agreement between the 4 raters at the 95% confidence level.

Table 1.4. One-way random effects ANOVA table relative to the example of Shrout and Fleiss (1979)

Variability	Sum of squares	Degrees of freedom	Mean squares
Between items	56.21	5	11.24
Within items	112.75	18	6.26
Total	168.96	23	

1.3.2 Two-way ANOVA models

Random raters (two-way random effects ANOVA model with interaction). Suppose now that each item is rated by the same random sample of R raters selected from a larger population (Model 2). The component $W_{i,r}$ can be further specified. A two-way model can be used to represent the data because there is a systematic source of variation between items and between raters. The component representing rater r effect may thus be estimated.

$$Y_{i,r} = \mu + B_i + A_r + (AB)_{i,r} + E_{i,r},\ (i = 1, \cdots, N; r = 1, \cdots, R). \tag{1.17}$$

The terms $Y_{i,r}$, μ and B_i were defined in Section 1.3.1. The component A_r denotes the deviation of rater r measurements from the overall mean, $(AB)_{i,r}$ is the degree to which the rater r departs from his/her usual rating tendencies when confronted to item i (interaction effect) and $E_{i,r}$ is the random error in ratings of rater r on item i. It is assumed that $A_r \sim N(0, \sigma_A^2)$, $B_i \sim N(0, \sigma_B^2)$, $E_{i,r} \sim N(0, \sigma_E^2)$ are independently distributed. Finally, all components $(AB)_{i,r}$ $(i = 1, \cdots, N; r = 1, \cdots, R)$ are assumed to be mutually independent and $(AB)_{i,r} \sim N(0, \sigma_I^2)$. The ANOVA table corresponding to Model 2 is given in Table 1.5.

Table 1.5. Two-way random effects ANOVA model with interaction (Model 2) (MS for mean squares)

Variability	Sum of squares	Degrees of freedom	MS	E(MS)
Between items	BSS	$N-1$	BMS	$R\sigma_B^2 + \sigma_I^2 + \sigma_E^2$
Within items	WSS	$N(R-1)$	WMS	$\sigma_A^2 + \sigma_I^2 + \sigma_E^2$
Between raters	JSS	$(R-1)$	JMS	$N\sigma_A^2 + \sigma_I^2 + \sigma_E^2$
Residuals	ESS	$(N-1)(R-1)$	EMS	$\sigma_I^2 + \sigma_E^2$
Total	TSS	$NR-1$		

Under Model 2, the intraclass correlation coefficient measuring absolute agreement (ICC_{A2}) and the consistency (ICC_{C2}) are defined by

$$ICC_{A2} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_A^2 + \sigma_I^2 + \sigma_E^2} \quad \text{and} \quad ICC_{C2} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_I^2 + \sigma_E^2}. \tag{1.18}$$

Fixed raters (two-way mixed effects ANOVA model). Model 3 is similar to Model 2 except that raters are considered as fixed.

$$Y_{i,r} = \mu + B_i + a_r + (aB)_{i,r} + E_{i,r},\ (i = 1, \cdots, N; r = 1, \cdots, R). \tag{1.19}$$

The same assumptions are made as in Model 2 for the components $Y_{i,r}$, μ and B_i but here, a_r is a fixed effect subject to the constraint $\sum_{r=1}^{R} a_r = 0$. The parameter

corresponding to σ_A^2 is

$$\theta_A^2 = \sum_{r=1}^{R} a_r^2/(R-1). \tag{1.20}$$

It is assumed that $(aB)_{i,r} \sim N(0, \sigma_I^2)$ $(i = 1, \cdots, N; r = 1, \cdots, R)$ but independence can only be assumed for interaction components that involve different items. For the same item i, the components are assumed to satisfy the constraint $\sum_{r=1}^{R}(aB)_{i,r} = 0$.

One implication of the raters being fixed is that no unbiased estimator of σ_B^2 is available when $\sigma_B^2 > 0$. σ_B^2 is no longer the covariance between $Y_{i,r}$ and $Y_{i,s}$ $(r \neq s)$. The interaction term has variance σ_I^2 and

$$cov(Y_{i,r}, Y_{i,s}) = \sigma_B^2 - \frac{\sigma_I^2}{R}. \tag{1.21}$$

The ANOVA table relative to Model 3 is given in Table 1.6 where $f = R/(R-1)$. It is crucial to note that the expectation of BMS under Models 2 and 3 is different of that under Model 1 even if the computation is the same. Because the effect of raters is the same for all items under Models 2 and 3, inter-rater variability does not affect the expectation of BMS. An important practical implication is that for a given population of items, the observed value of BMS in a Model 1 design tends to be larger than in a Model 2 or 3 design.

Table 1.6. Two-way mixed effects ANOVA model (Model 3) (MS for mean squares)

Variability	Sum of squares	Degrees of freedom	MS	E(MS)
Between items	BSS	$N-1$	BMS	$R\sigma_B^2 + \sigma_E^2$
Within items	WSS	$N(R-1)$	WMS	$\theta_A^2 + f\sigma_I^2 + \sigma_E^2$
Between raters	JSS	$(R-1)$	JMS	$N\theta_A^2 + f\sigma_I^2 + \sigma_E^2$
Residuals	ESS	$(N-1)(R-1)$	EMS	$f\sigma_I^2 + \sigma_E^2$
Total	TSS	$NR-1$		

Under Model 3, the intraclass correlation coefficients ICC_{A3} and ICC_{C3} are defined as

$$ICC_{A3} = \frac{\sigma_B^2 - \sigma_I^2/(R-1)}{\sigma_B^2 + \theta_A^2 + \sigma_I^2 + \sigma_E^2} \quad \text{and} \quad ICC_{C3} = \frac{\sigma_B^2 - \sigma_I^2/(R-1)}{\sigma_B^2 + \sigma_I^2 + \sigma_E^2}. \tag{1.22}$$

Estimation of the parameters. Although the definition of the agreement indexes are different depending if raters are considered as random or fixed, the estimated intraclass correlations are the same ($\widehat{ICC}_{A2} = \widehat{ICC}_{A3} = \widehat{ICC}_A$ and $\widehat{ICC}_{C2} = \widehat{ICC}_{C3} = \widehat{ICC}_C$). The intraclass correlation coefficients are estimated

by

$$\widehat{ICC}_A = \frac{BMS - EMS}{BMS + (R-1)EMS + R(JMS - EMS)/N} \tag{1.23}$$

and

$$\widehat{ICC}_C = \frac{BMS - EMS}{BMS + (R-1)EMS}. \tag{1.24}$$

The agreement coefficient $\widehat{ICC}_{A2}$ is also known as the *criterion-referenced reliability* and the agreement coefficient $\widehat{ICC}_{C2}$ as *norm-referenced reliability* and as *Winer's adjustment for anchor points* (McGraw and Wong, 1996).

The quantities $\overline{y}_{i,\cdot}$ and $\overline{y}_{\cdot,\cdot}$ are defined as previously. Let $\overline{y}_{\cdot,r}$ denote the mean over the items for rater r ($r = 1, \cdots, R$).

$$\overline{y}_{\cdot,r} = \frac{1}{N}\sum_{i=1}^{N} y_{i,r}.$$

We have

$$\begin{aligned}
BSS &= R\sum_{i=1}^{N}(\overline{y}_{i,\cdot} - \overline{y}_{\cdot,\cdot})^2, \\
JSS &= N\sum_{r=1}^{R}(\overline{y}_{\cdot,r} - \overline{y}_{\cdot,\cdot})^2, \\
ESS &= \sum_{i=1}^{N}\sum_{r=1}^{R}(y_{i,r} - \overline{y}_{i,\cdot} - \overline{y}_{\cdot,r} + \overline{y}_{\cdot,\cdot})^2, \\
TSS &= \sum_{i=1}^{N}\sum_{r=1}^{R}(y_{i,r} - \overline{y}_{\cdot,\cdot})^2.
\end{aligned} \tag{1.25}$$

Confidence interval for ICC_A. Let $F_0 = BMS/EMS$ be the usual variance ratio distributed as a Snedecor F with $N-1$ and $(N-1)(R-1)$ degrees of freedom. The confidence interval is more complicated to derive since the index is a function of three independent mean squares. Following Satterwhaite (1946), Fleiss and Shrout (1978) derived an approximate confidence interval. Let $F_J = JMS/EMS$ and

$$\nu = \frac{(R-1)(N-1)\{R\ \widehat{ICC}_A F_J + N(1 + (R-1)\widehat{ICC}_A) - R\ \widehat{ICC}_A\}^2}{(N-1)R^2\widehat{ICC}_A^2 F_J^2 + \{N(1 + (R-1)\widehat{ICC}_A) - R\ \widehat{ICC}_A\}^2}. \tag{1.26}$$

The lower bound of the $(1-\alpha)100\%$ confidence interval for ICC_A is defined by

$$\frac{N(BMS - F_U\ EMS)}{F_U R\ JMS + (RN - R - N)EMS + N\ BMS} \tag{1.27}$$

and the upper bound by

$$\frac{N(F_L BMS - EMS)}{R\ JMS + (RN - R - N)EMS + N\ F_L BMS}. \tag{1.28}$$

where

$$\begin{aligned} F_L &= Q_F(1 - \frac{\alpha}{2}; \nu, N-1) \\ F_U &= Q_F(1 - \frac{\alpha}{2}; N-1, \nu). \end{aligned}$$

Confidence interval for ICC_C. If $F_0 = BMS/EMS$ denotes the usual variance ratio distributed as a Snedecor F with $N-1$ and $(N-1)(R-1)$ degrees of freedom,

$$\frac{F_L - 1}{F_L + (R-1)} < ICC_C < \frac{F_U - 1}{F_U + (R-1)} \tag{1.29}$$

is a $(1-\alpha)100\%$ confidence interval for ICC_C with

$$\begin{aligned} F_L &= F_0 / Q_F(1 - \frac{\alpha}{2}; N-1, (N-1)(R-1)) \\ F_U &= F_0 Q_F(1 - \frac{\alpha}{2}; (N-1)(R-1), N-1). \end{aligned}$$

Example. Consider again the example of Table 1.1 but suppose now that the same set of 4 raters have all measured the 6 items on a 10-point scale and that the 4 raters are taken at random from a larger population of raters. This leads to the following ANOVA table (Table 1.7).

Table 1.7. Two-way random effects ANOVA model with interaction (Model 2) relative to the example of Shrout and Fleiss (1979)

Variability	Sum of squares	Degrees of freedom	Mean squares
Between items	56.21	5	11.24
Within items	112.75	18	6.26
Between raters	97.46	3	32.49
Residuals	15.29	15	1.02
Total	168.96	23	

The intraclass correlation coefficient with the absolute definition is then estimated by

$$\widehat{ICC}_{A2} = \frac{11.24 - 1.02}{11.24 + 3 \times 1.02 + 4(32.49 - 1.02)/6} = 0.29.$$

We have $\nu = 33123.31/6922.11 = 4.79$, $F_0 = 11.24/1.02 = 11.02$ and $F_J = 32.49/1.02 = 31.85$. The one-sided lower bound at 95% confidence level is equal

to 0.05, indicating that there is a slight agreement between the 4 raters. We also have

$$\widehat{ICC}_{C2} = \frac{11.24 - 1.02}{11.24 + 3 \times 1.02} = 0.71$$

with 95% one-sided lower bound of 0.41 where $F_L = 11.03/2.90 = 3.80$, meaning that the measurements of the 4 raters are consistent.

1.3.3 Mean of individual ratings

The ICCs discussed before express the expected reliability of the measurements of single raters. Sometimes, it is not the individual ratings that are used but rather the mean of m ratings ($m \leq R$), i.e. m is not necessarily equal to the number of raters in the study. The unit of analysis is then a mean of ratings rather than individual ratings. In such case the reliability of the mean rating is of interest; the reliability will always be greater in magnitude than the reliability of the individual ratings, provided the latter is positive (Lord and Novick, 1968). An example of a substantive choice is the investigation of the decisions (ratings) of a team of physicians, as they are found in a hospital setting. More typically, an investigator decides to use a mean as a unit of the analysis because the individual ratings are too unreliable (Shrout and Fleiss, 1979). The number of raters (i.e., m) used to form the mean ratings needs to be determined. Given a lower bound, ICC_L, and the minimum acceptable value of the reliability coefficient ICC^* (e.g., $ICC^* = 0.75$) it is possible to determine m as the smallest integer greater than or equal to (Shrout and Fleiss, 1979)

$$m = \frac{ICC^*(1 - ICC_L)}{ICC_L(1 - ICC^*)}. \tag{1.30}$$

Once m is determined, either by a reliability study or by a choice made on substantive grounds, the reliability of the ratings averaged over the m raters can be estimated using the appropriate intraclass correlation coefficient described earlier. When data from m raters are actually collected, they can be used to estimate the reliability of the mean ratings in one step, using the formulas below, depending of the study design. In these applications, we suppose that $m = R$.

One-way random effects ANOVA model. The intraclass correlation coefficient corresponding to the one-way random effects ANOVA model when considering averaged measurements rather than single measurements is defined by

$$ICC_{A1,R} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2/R} \tag{1.31}$$

and is estimated by

$$\widehat{ICC}_{A1,R} = \frac{BMS - WMS}{BMS}. \tag{1.32}$$

Letting F_U and F_L defined as for ICC_{A1},

$$1 - \frac{1}{F_L} < ICC_{A1,R} < 1 - \frac{1}{F_U} \tag{1.33}$$

is a $(1-\alpha)100\%$ confidence interval for $ICC_{A1,R}$.

Two-way random effects ANOVA model with interaction. The degree of absolute agreement is expressed as

$$ICC_{A2,R} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_A^2 + (\sigma_I^2 + \sigma_E^2)/R} \tag{1.34}$$

and estimated by

$$\widehat{ICC}_{A2,R} = \frac{BMS - EMS}{BMS + (JMS - EMS)/N}. \tag{1.35}$$

The confidence interval used the confidence bounds obtained for ICC_{A2}. For example, the lower bound for $ICC_{2,R}$ is

$$ICC_L = \frac{RICC_L^{**}}{1 + (R-1)ICC_L^{**}} \tag{1.36}$$

where ICC_L^{**} is the lower bound obtained for ICC_{A2}.

On the other hand, the degree of consistency can be quantified using

$$ICC_{C2,R} = \frac{\sigma_B^2}{\sigma_B^2 + (\sigma_I^2 + \sigma_E^2)/R} \tag{1.37}$$

and estimated by

$$\widehat{ICC}_{C2,R} = \frac{BMS - EMS}{BMS}. \tag{1.38}$$

Let F_U and F_L be defined as for ICC_{C2},

$$1 - \frac{1}{F_L} < ICC_{C2,R} < 1 - \frac{1}{F_U} \tag{1.39}$$

is a $(1-\alpha)100\%$ confidence interval for $ICC_{C2,R}$. Note that $ICC_{C2,R}$ is equivalent to *Cronbach's alpha* (Cronbach, 1951). Cronbach's alpha will generally increase when the correlation between the items increases. For this reason the coefficient is also called the *internal consistency* or the *internal consistency reliability* of the test (Shrout and Fleiss, 1979). Internal consistency is a measure based on the

correlations between different items on the same test (or the same subscale on a larger test). It measures whether several items proposed to measure the same general construct produce similar scores. For example, if a respondent expressed agreement with the statements "I like to ride bicycles" and "I've enjoyed riding bicycles in the past", and disagreement with the statement "I hate bicycles", this would be indicative of good internal consistency of the test. A commonly-accepted rule of thumb is that a Cronbach's alpha coefficient of 0.6-0.7 indicates acceptable reliability, and 0.8 or higher indicates good reliability. Note that extremely high reliabilities (0.95 or higher) are not necessarily desirable, indicating that the items may be not just consistent, but redundant.

Alternatively, Cronbach's alpha coefficient can also be defined as

$$\widehat{ICC}_{C2,R} = \frac{N\overline{\hat{\rho}}}{1+(N-1)\overline{\hat{\rho}}} \tag{1.40}$$

where $\overline{\hat{\rho}}$ is the average of all Pearson's correlation coefficients between the items.

Two-way mixed effects ANOVA model. The generalization from single rating to mean rating reliability is not quite as straightforward as in the random effects model. Although the covariance between two ratings is $\sigma_B^2 - \sigma_I^2/(R-1)$, the covariance between two means based on R raters is σ_B^2. No estimator exists for this term. If, however, the rater $\times$ item interaction can be assumed to be absent ($\sigma_I^2 = 0$), the agreement indexes $ICC_{A3,R}$ and $ICC_{C3,R}$ are defined by

$$ICC_{A3,R} = \frac{\sigma_B^2 - \sigma_I^2/(R-1)}{\sigma_B^2 + (\theta_A^2 + \sigma_I^2 + \sigma_E^2)/R} = \frac{\sigma_B^2}{\sigma_B^2 + (\theta_A^2 + \sigma_E^2)/R} \tag{1.41}$$

and

$$ICC_{C3,R} = \frac{\sigma_B^2 - \sigma_I^2/(R-1)}{\sigma_B^2 + (\sigma_I^2 + \sigma_E^2)/R} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_E^2/R} \tag{1.42}$$

and estimated by $\widehat{ICC}_{A2,R}$ and $\widehat{ICC}_{C2,R}$, respectively.

Example. Consider again the example described in Table 1.1 and suppose that measurements are average measures rather than single measures and let determine Cronbach's alpha coefficient of reliability. We have $\widehat{ICC}_{C2,4} = (11.24 - 1.02)/11.24 = 0.91$ (see Table 1.7). We obtained, in case of individual measurements $F_L = 3.80$, leading to a one-sided lower bound at 95% confidence level of $1 - 1/3.80 = 0.74$. There is thus a good consistency between the raters.

1.4 Serum gentamicin

Serum gentamicin ($\mu mol/L$) was measured by two assay methods, the enzyme-mediated immunoassay technique (EMIT), used in routine at the time of the study,

and the fluoro-immunoassay (FIA), a new method to be tested (Strike, 1991). Data are given in Appendix A (see Table A.1). Serum specimens from 56 patients receiving gentamicin have been assayed twice by each assay method in separate assay batches. Agreement between the two methods is needed to validate the new assay method. Firstly, the mean value of the two repeated measurements was calculated for each method. The resulting 45° line plot and Bland and Altman plot are given in Figure 1.2. The mean of the differences between the two methods of measurements was equal to -0.084 ± 1.18 with 95%CI: [-2.39,2.22]. Thus, there was no systematic bias between the two methods.

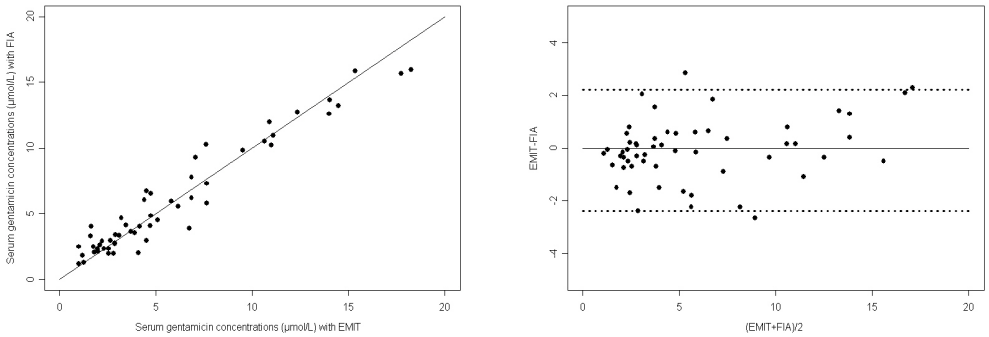

Figure 1.2. Serum gentamicin concentrations ($\mu mol/L$) measured with the EMIT and the FIA methods on a 45° line plot (left) and Bland and Altman plot (right) with 95% confidence interval

The concordance correlation coefficient was equal to $\widehat{CCC} = 0.96$ with a lower bound of two-sided 95% CI equal to 0.90. The two-way mixed effects ANOVA table corresponding to the 4 separate measurements is given in Table 1.8.

Table 1.8. Two-way mixed effects ANOVA model (Model 3)

Variability	Sum of squares	Degrees of freedom	Mean squares
Between patients	4105.72	55	74.65
Between methods	4.00	3	1.33
Residuals	129.71	165	0.79
Total	4239.43	223	

Assuming no interaction between the methods and the patients, the intraclass correlation coefficient using the definition of absolute agreement was equal to $\widehat{ICC}_{A3,2} = 0.99$ with a one-sided 95% lower confidence bound of 0.97. The intraclass correlation coefficient between the two methods using the consistency definition was equal to $\widehat{ICC}_{C3,2} = 0.99$ with a one-sided 95% lower bound of 0.99.

This indicated quite good agreement between the two methods, suggesting that the FIA method could be used confidently in daily routine.

1.5 Discussion

Some basic statistical approaches for assessing agreement between two or more raters on a quantitative scale were reviewed in this chapter. This is common practice when validating new measurement methods with respect to existing ones. When comparing two methods, one method is usually the reference method and the other method a new method to be tested. The problem is therefore one with fixed raters rather than random raters (calibration problem).

The simplest practical way to assess the agreement between two raters on a quantitative scale is from the visual plots (deviation from the 45° line or Bland and Altman plot). The difficulty in Bland and Altman plot is to determine what is a reasonable 95% CI around the mean difference between the two measurements. Therefore, the need for formal testing and quantification of the amount of agreement between raters led Lin (1989) to define the concordance correlation coefficient (CCC). In recent years, the CCC was extended to various situations (Barnhart and Williamson, 2002; King and Chinchilli, 2001; Lin et al., 2002).

The intraclass correlation coefficients (ICC) are also widely used to quantify agreement on quantitative scales and should be used with care depending on the study design and the question of interest. As stated by Shrout and Fleiss (1979), an important issue is the choice of an appropriate ANOVA model. This was seen with the example of Shrout and Fleiss (1979), where conclusions were different depending on the ANOVA model chosen. McGraw and Wong (1996) reviewed intraclass correlation coefficients introduced by Shrout and Fleiss (1979) and made the distinction between accuracy and absolute agreement indexes. They also stressed the difference between treating the raters as fixed or random. In practical terms, one knows that the levels of a variable are random when a change in those levels of the variable would have no effect on the question being asked. As an example of fixed effect variables, McGraw and Wong (1996) considered the biological relation between mother and child. By changing the levels in uncle and nephew would imply a totally different research interest. Although the estimated intraclass correlation coefficients are the same when raters are treated as fixed or as random, the interpretation is different. When treated as random, the results can be generalized at the population level which is not the case when treated as fixed. Carrasco and Jover (2003) showed that the CCC is equivalent to the ICC under a two-way mixed effects ANOVA model with fixed raters. A more detailed review of methods to quantify agreement on a quantitative scale is given in Barnhart et al. (2007).

CHAPTER 2

Agreement between two independent raters

2.1 Introduction

The problem of rater agreement on a categorical scale originally emerged in human sciences, where measurements are 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. Medicine is not an exact science but we would expect that physicians tend to agree with each other. The validation process of any new scale also requires the study of agreement among raters. The simplest case is to determine the agreement between two raters (methods or observers) on a binary scale (e.g. diseased/ non diseased). Several coefficients for quantifying the agreement between two raters have been introduced over the years. The most salient one is the kappa coefficient introduced by Cohen (1960). It is the most widely used coefficient of agreement in scientific research (Blackman and Koval, 2000; Ludbrook, 2002). Cohen's kappa coefficient differs from the others in the sense that it accounts for agreement between the two raters due to chance. Indeed, if two raters randomly assign a series of items on a categorical scale, the observed agreement between them is then only due to chance. Cohen (1968) also introduced the weighted kappa coefficient to allow for the fact that some disagreements may be more important than others. Indeed, disagreements between two raters occurring on the categories "light" and "severe"

may be viewed as more important than on "light" and "moderate". Finally, Kraemer (1979) defined a kappa coefficient by assuming that the two raters have the same marginal distribution. Agreement indexes are reviewed in this chapter and their asymptotic sampling variance derived.

2.2 Early agreement indexes

Consider two independent raters who have to classify a sample of N items (subjects or objects) into K exhaustive and mutually exclusive categories of a nominal or ordinal scale. The observations made by the 2 raters can be summarized in a $K \times K$ contingency table (Table 2.1), where n_{jk} is the number of items classified in category j by rater 1 and category k by rater 2; let $n_{j.}$ be the number of items classified in category j by rater 1 and $n_{.k}$ the number of items classified in category k by rater 2. By dividing these numbers by N, the corresponding proportions p_{jk}, $p_{j.}$, $p_{.k}$ are obtained.

Table 2.1. $K \times K$ contingency table summarizing the classification of N items by 2 raters on a K-category scale in terms of frequency (proportion)

	Rater 2					
Rater 1	1	...	j	...	K	Total
1	n_{11} (p_{11})	$\cdots$	n_{1j} (p_{1j})	...	n_{1K} (p_{1K})	$n_{1.}$ $(p_{1.})$
$\vdots$	$\vdots$		$\vdots$	$\vdots$		
j	n_{j1} (p_{j1})	$\cdots$	n_{jj} (p_{jj})	...	n_{jK} (p_{jK})	$n_{j.}$ $(p_{j.})$
$\vdots$	$\vdots$		$\vdots$	$\vdots$		
K	n_{K1} (p_{K1})	$\cdots$	n_{Kj} (p_{Kj})	...	n_{KK} (p_{KK})	$n_{K.}$ $(p_{K.})$
Total	$n_{.1}$ $(p_{.1})$	$\cdots$	$n_{.j}$ $(p_{.j})$	...	$n_{.K}$ $(p_{.K})$	N (1)

When there are only two categories (binary case), Table 2.1 reduces to a 2×2 contingency table, where the categories are often labeled as 0 and 1 (see Table 2.2).

Intuitively, it seems obvious to use the sum of the diagonal proportions of Table 2.2 to quantify the agreement between the two raters. Indeed, it represents the proportion of items classified in the same category by the 2 raters. It is called the *observed proportion of agreement*

$$p_o = p_{11} + p_{22}.$$

The index p_o is the simplest agreement index (Holley and Guilford, 1964; Maxwell, 1977).

Table 2.2. 2×2 contingency table corresponding to the classification of N items on a binary scale by 2 raters in terms of frequency (proportion)

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

Suppose that the trait under study is relatively rare. In that case, negative agreements (p_{22}) may be more frequent than positive agreements (p_{11}). Then, it may be reasonable to omit the proportion p_{22} in the construction of the agreement index because that proportion will be large (since the trait under study is rare) and inflate the value p_o. For that reason, a number of indexes based only on the proportions p_{11}, p_{12} and p_{21} were proposed. The index proposed by Dice (1945) was

$$S_d = \frac{p_{11}}{\frac{1}{2}(p_{1.} + p_{.1})}.$$

The index S_d can be interpreted as a conditional probability. Indeed, if we randomly choose one of the two raters and consider the items classified positive by this rater, S_d is the conditional probability that the second rater classifies the item positive while the first rater classified the item positive. The same index exists if we decide to ignore the proportion p_{11} instead of the proportion p_{22}. This index writes

$$S'_d = \frac{p_{22}}{\frac{1}{2}(p_{2.} + p_{.2})}.$$

Rogot and Goldberg (1966) proposed to take the mean of S_d et S'_d as agreement index between the two raters

$$A_2 = \frac{p_{11}}{(p_{1.} + p_{.1})} + \frac{p_{22}}{(p_{2.} + p_{.2})}.$$

Note that $A_2 = 1$ in case of perfect agreement. Goodman and Kruskal (1972) suggested the following index

$$\lambda_r = \frac{2p_{11} - (p_{12} + p_{21})}{2p_{11} + (p_{12} + p_{21})}.$$

It is easily seen that $\lambda_r = 2S_d - 1$. The maximum value of λ_r is 1 when agreement is perfect and the minimum value is -1 when $p_{11} = 0$.

In all coefficients given above, the agreements due to chance alone have not been taken in account. However, Scott (1955) introduced an index of inter-rater agreement taking into account chance agreement. This is known as Scott's π,

$$\hat{\pi} = \frac{p_o - p_e}{1 - p_e} \tag{2.1}$$

where p_o is defined as earlier and p_e is the proportion of agreement to be expected by chance, namely,

$$p_e = p_1^2 + p_2^2. \tag{2.2}$$

In this expression, $p_j = (p_{.j} + p_{j.})/2$ is the overall proportion of items in the sample classified in category j ($j = 1, 2$). A more general way to correct for chance effect is introduced in the next section.

2.3 Cohen's kappa coefficient

2.3.1 Binary scale

Cohen (1960) introduced two proportions to define an agreement index between two independent raters on a binary scale, the *observed proportion of agreement*

$$p_o = \frac{n_{11} + n_{22}}{N} = p_{11} + p_{22} \tag{2.3}$$

and the *proportion of agreement expected by chance*

$$p_e = \frac{n_{1.}n_{.1} + n_{2.}n_{.2}}{N^2} = p_{1.}p_{.1} + p_{2.}p_{.2}. \tag{2.4}$$

To define the agreement index, Cohen (1960) considered the observed proportion of agreement after that the proportion of agreement expected by chance is removed from consideration. The result is then scaled to obtain a value 1 when agreement is perfect, a value 0 when agreement is only due to chance and negative values when observed agreement is lower than agreement expected by chance. Specifically, Cohen's kappa coefficient writes

$$\widehat{\kappa} = \frac{p_o - p_e}{1 - p_e}. \tag{2.5}$$

2.3.2 Categorical scale

By extension, Cohen (1960) defined the *observed proportion of agreement* on a categorical scale by

$$p_o = \sum_{j=1}^{K} \frac{n_{jj}}{N} = \sum_{j=1}^{K} p_{jj} \tag{2.6}$$

and the *proportion of agreement expected by chance* by

$$p_e = \sum_{j=1}^{K} \frac{n_{j.} n_{.j}}{N^2} = \sum_{j=1}^{K} p_{j.} p_{.j}, \tag{2.7}$$

leading to the Cohen's kappa coefficient

$$\widehat{\kappa} = \frac{p_o - p_e}{1 - p_e}. \tag{2.8}$$

Landis and Koch (1977b) proposed to qualify the strength of agreement according to the values taken by Cohen's kappa coefficient (see Table 2.3), although no longer recommended today because the divisions are clearly arbitrary and vary depending on the problem under study. The precision with which Cohen's kappa coefficient is estimated is also an important aspect (statistical significance).

Table 2.3. Qualification of the strength of agreement according to values of $\hat{\kappa}$ following Landis and Koch (1977b)

Agreement	$\hat{\kappa}$
Almost perfect	> 0.81
Substantial	$0.61 - 0.80$
Moderate	$0.41 - 0.60$
Fair	$0.21 - 0.40$
Slight	$0 - 0.20$
Poor	< 0

2.3.3 Properties

Hereafter, we look at the properties of Cohen's kappa coefficient.

Property 1. $\hat{\kappa} = 1$ *if and only if* $p_{ij} = 0$ *(*$i \neq j \in 1, \cdots, K$*).*

The upper limit of $\widehat{\kappa}$ is equal to 1, occurring if and only if there is perfect agreement between the two raters. The condition : $p_{.j} = p_{j.},\ j = 1, \ldots, K$ is necessary but not sufficient to have perfect agreement. Indeed, if $\exists j \in 1, \cdots, K : p_{.j} \neq p_{j.}$, there is automatically disagreement.

Property 2. *Given fixed margins, the maximum value of* $\hat{\kappa}$ *is obtained for* $p_o = p_{oM} = \sum_{j=1}^{K} inf(p_{j.}, p_{.j})$.

For given margins, Cohen (1960) proposed the following expression, in order to determine the maximum value of $\widehat{\kappa}$:

$$\hat{\kappa}_M = \frac{p_{oM} - p_e}{1 - p_e} \tag{2.9}$$

where

$$p_{oM} = \sum_{j=1}^{K} inf(p_{j.}, p_{.j})$$

is the maximum proportion of observed agreement permitted by the marginals.

Property 3. *The minimum value of $\hat{\kappa}$ is obtained for $p_o = 0$.*

The lower limit $\hat{\kappa}_m$ of Cohen's kappa coefficient is attained when the observed proportion of agreement p_o between the two raters is nil. Thus,

$$\hat{\kappa}_m = -\frac{p_e}{1 - p_e}. \tag{2.10}$$

The lower limit $\hat{\kappa}_m$ only depends on the marginal distributions $p_{j.}$ and $p_{.j}$, $(j = 1, \cdots, K)$ since it only involves the proportion p_e and depends on the direction of the two ratings. If the ratings go in the same direction, then $\hat{\kappa}_m < -1/(K-1)$. Otherwise, $\hat{\kappa}_m \geq -1$. When the scale is binary, $\hat{\kappa}_m \geq -1$.

Property 4. *Relationship between binary and K-category scales*

Cohen's kappa coefficient relative to a K-category scale can be derived from Cohen's kappa coefficients derived on a binary scale obtained by isolating a category j from the other categories $(j = 1, \cdots, K)$. This leads to the contingency table displayed in Table 2.4.

Table 2.4. 2×2 contingency table obtained by isolating category j from the other categories $(j = 1, \cdots, K)$ in terms of frequency (proportion)

	Rater 2		
Rater 1	Category j	Other categories	Total
Category j	n_{jj}	$n_{j.} - n_{jj}$	$n_{j.}$
Other categories	$n_{.j} - n_{jj}$	$N - n_{.j} - n_{j.} + n_{jj}$	$N - n_{j.}$
Total	$n_{.j}$	$N - n_{.j}$	N

From this contingency table, the proportion of observed agreement is defined by

$$p_{o[j]} = \frac{n_{jj} + N - n_{.j} - n_{j.} + n_{jj}}{N} = p_{jj} + 1 - p_{.j} - p_{j.} + p_{jj} \tag{2.11}$$

and the proportion of agreement expected by chance relative to category j by

$$p_{e[j]} = \frac{n_{j.}n_{.j} + (N - n_{j.})(N - n_{.j})}{N^2} = p_{j.}p_{.j} + (1 - p_{j.})(1 - p_{.j}) \tag{2.12}$$

leading to the Cohen's kappa coefficient relative to category j,

$$\hat{\kappa}_{[j]} = \frac{p_{o[j]} - p_{e[j]}}{1 - p_{e[j]}}. \tag{2.13}$$

Remark that agreement (diagonal) and disagreement (off-diagonal) cells of the $K \times K$ contingency table are mixed in the quantity $N - n_{.j} - n_{j.} + n_{jj}$. The overall Cohen's kappa coefficient can then be rewritten

$$\widehat{\kappa} = \frac{\sum_{j=1}^{K}(p_{o[j]} - p_{e[j]})}{\sum_{j=1}^{K}(1 - p_{e[j]})} = \frac{1}{\sum_{j=1}^{K}(1 - p_{e[j]})} \sum_{j=1}^{K}(1 - p_{e[j]})\widehat{\kappa}_{[j]}. \tag{2.14}$$

Property 5. *Relation between $\hat{\kappa}$ and Pearson's chi-square $\hat{\phi}$ coefficients for binary scales*

Cohen (1960) investigated the relation between Cohen's kappa ($\hat{\kappa}$) and Pearson's chi-square ($\hat{\phi}$) coefficients for binary scales. The coefficient $\widehat{\phi}$ can be expressed as followed :

$$\widehat{\phi} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{.1}n_{.2}n_{1.}n_{2.}}}. \tag{2.15}$$

By simple algebraic transformations, Cohen's kappa coefficient can be written as followed:

$$\widehat{\kappa} = \frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{.1}n_{2.} + n_{.2}n_{1.}}. \tag{2.16}$$

If we suppose that $n_{.1} = vN$ ($n_{.2} = (1 - v)N$) and $n_{1.} = wN$ ($n_{2.} = (1 - w)N$), with $0 \leq v, w \leq 1$, then

$$\widehat{\phi}^2 = \widehat{\kappa}^2 \left(1 + \frac{N^4(v - w)^2}{4n_{.1}n_{.2}n_{1.}n_{2.}}\right). \tag{2.17}$$

In general, $\widehat{\phi}^2 \geq \widehat{\kappa}^2$, $\widehat{\phi}^2 = \widehat{\kappa}^2$ if and only if $n_{.1} = n_{.2} = n_{1.} = n_{2.} = N/2$.

Property 6. *Population model and maximum likelihood estimator*

Cohen (1960) defined Cohen's kappa coefficient as a descriptive statistic on an ad hoc basis and not in terms of population parameters. However, Bloch and Kraemer (1989) derived a population model in the case of a binary scale yielding the Cohen's kappa coefficient as maximum likelihood estimator.

Consider a population of items $\mathcal{I}$. Let $Y_{i,r}$ be the random variable such that $Y_{i,r} = 1$ if rater r $(r = 1, 2)$ classifies a randomly selected item i of population $\mathcal{I}$ in category 1 and $Y_{i,r} = 0$ otherwise. Over the population of items, $E(Y_{i,r}) = \pi_r$ and $var(Y_{i,r}) = \sigma_r^2 = \pi_r(1 - \pi_r)$. If ρ denotes the correlation between $Y_{i,1}$ and $Y_{i,2}$, Table 2.5 corresponds to the population model.

Table 2.5. Theoretical model in the case of two independent raters and a binary scale

	Rater 2		
Rater 1	0	1	
0	$E[(1 - Y_{i,1})(1 - Y_{i,2})]$ $(1 - \pi_1)(1 - \pi_2) + \rho\sigma_1\sigma_2$	$E[(1 - Y_{i,1})Y_{i,2}]$ $(1 - \pi_1)\pi_2 - \rho\sigma_1\sigma_2$	$1 - \pi_1$
1	$E[Y_{i,1}(1 - Y_{i,2})]$ $\pi_1(1 - \pi_2) - \rho\sigma_1\sigma_2$	$E[Y_{i,1}Y_{i,2}]$ $\pi_1\pi_2 + \rho\sigma_1\sigma_2$	π_1
	$1 - \pi_2$	π_2	1

Cohen's kappa coefficient is then defined as

$$\begin{aligned}
\kappa &= \frac{\text{Expected agreement} - \text{Random agreement}}{\text{Maximum expected agreement} - \text{Random agreement}} \\
&= \frac{[\pi_1\pi_2 + (1 - \pi_1)(1 - \pi_2) + 2\rho\sigma_1\sigma_2] - [\pi_1\pi_2 + (1 - \pi_1)(1 - \pi_2)]}{1 - \pi_1\pi_2 + (1 - \pi_1)(1 - \pi_2)} \\
&= \frac{2\rho\sigma_1\sigma_2}{1 - \pi_1\pi_2 + (1 - \pi_1)(1 - \pi_2)}. \qquad (2.18)
\end{aligned}$$

Suppose that the two raters classify a random sample of N items from population $\mathcal{I}$ on a binary scale. This leads to the contingency table displayed in Table 2.2. The log-likelihood function is then

$$\begin{aligned}
&\ln L(\pi_1, \pi_2, \kappa | n_{11}, n_{12}, n_{21}, n_{22}) \\
&+n_{11} \ln[\pi_1\pi_2 + \frac{1}{2}\kappa(\pi_1(1 - \pi_2) + (1 - \pi_1)\pi_2)] \\
&+n_{12} \ln[(\pi_1(1 - \pi_2) - \frac{1}{2}\kappa(\pi_1(1 - \pi_2) + (1 - \pi_1)\pi_2)] \\
&+n_{21} \ln[(1 - \pi_1)\pi_2 - \frac{1}{2}\kappa(\pi_1(1 - \pi_2) + (1 - \pi_1)\pi_2)] \\
&= n_{22} \ln[(1 - \pi_1)(1 - \pi_2) + \frac{1}{2}\kappa(\pi_1(1 - \pi_2) + (1 - \pi_1)\pi_2)]. \qquad (2.19)
\end{aligned}$$

The maximum likelihood estimators of π_1, π_2 and κ are respectively

$$\hat{\pi}_1 = p_{1.} = \frac{n_{11} + n_{12}}{N} = p_{11} + p_{12}, \tag{2.20}$$

$$\hat{\pi}_2 = p_{.1} = \frac{n_{11} + n_{21}}{N} = p_{11} + p_{21}, \tag{2.21}$$

and

$$\widehat{\kappa} = \frac{2(n_{11}n_{22} - n_{12}n_{21})}{\hat{\pi}_1(1-\hat{\pi}_2) + \hat{\pi}_2(1-\hat{\pi}_1)} = \frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{.1}n_{2.} + n_{.2}n_{1.}}, \tag{2.22}$$

corresponding to Equation 2.16 and thus to the original definition given by Cohen (1960).

2.3.4 Sampling variability

Delta method. The expression of the large sample variance of Cohen's kappa coefficient given by the Delta method (Bishop et al., 1975) is

$$var(\hat{\kappa}) = \frac{p_o(1-p_o)}{N(1-p_e)^2} + \frac{2(p_o-1)(C_1 - 2p_o p_e)}{N(1-p_e)^3} + \frac{(p_o-1)^2(C_2 - 4p_e^2)}{N(1-p_e)^4} \tag{2.23}$$

where

$$C_1 = \sum_{j=1}^{K} p_{jj}(p_{j.} + p_{.j}) \text{ and } C_2 = \sum_{j=1}^{K}\sum_{k=1}^{K} p_{jk}(p_{.j} + p_{k.})^2.$$

The Delta method is exposed in Appendix B in the general case and in the particular case of multinomial data.

Garner's method. Garner (1991) derived a simple expression for an approximate large sample variance of Cohen's kappa for binary scales and a general procedure for K-categorical scales. Only the binary case will be exposed here. Garner (1991) took the following theoretical representation of the ratings made by 2 raters (Table 2.6). Note the similarity with Table 2.5 when considering $\delta = \rho\sigma_1\sigma_2$.

Table 2.6. Garner's theoretical representation of ratings made by two raters on a binary scale

		Rater 2		
		0	1	Total
Rater 1	0	$(1-\pi_1)(1-\pi_2)+\delta$	$(1-\pi_1)\pi_2 - \delta$	$1-\pi_1$
	1	$\pi_1(1-\pi_2) - \delta$	$\pi_1\pi_2 + \delta$	π_1
Total		$1-\pi_2$	π_2	1

Cohen's kappa coefficient is directly related to δ through the formula

$$\kappa = \frac{2\delta}{1 - [\pi_1\pi_2 + (1-\pi_1)(1-\pi_2)]}. \tag{2.24}$$

Conditioning on the observed marginal values, the theoretical proportion in the (1,1) cell is $\pi_1\pi_2+\delta$ and the observed proportion is $\hat{\pi}_1\hat{\pi}_2+\hat{\delta}$. Therefore the difference between the two is $\pm(\delta-\hat{\delta})$ in each cell. Garner (1991) used the fact that, in large samples, the conditional log-likelihood may be approximated by $(-1/2)\chi^2$, where χ^2 may be taken as the following sum over the four cells:

$$(observed\ frequency - expected\ frequency)^2/(an\ estimate\ of\ the\ cell\ frequency).$$

If $(\delta-\hat{\delta})$ denotes the difference between an observed and expected cell proportion, the large sample approximation to twice the negative log-likelihood may be written as

$$\chi^2 = \{N(\delta-\hat{\delta})\}^2 \left[\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{1}{Np^*_{jk}}\right]$$

where Np^*_{jk} is the observed cell frequency or some 'smoothed' estimate thereof. Since χ^2 has an asymptotically chi-square distribution with one degree of freedom when the four cell frequencies are 'large',

$$\chi^2 \approx \left(\frac{\delta-\hat{\delta}}{SE(\hat{\delta})}\right)^2 \text{ where } SE(\hat{\delta}) \approx \frac{1}{N\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{1}{Np^*_{jk}}}.$$

This yields the following large sample variance estimate for $\hat{\kappa}$,

$$var(\hat{\kappa}) = \frac{4}{(1-p_e)^2N^2\left[\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{1}{Np^*_{jk}}\right]^2}. \tag{2.25}$$

Garner (1991) proposed to replace Np^*_{jk} by $n_{jk}+1$ to avoid the problem of having a zero cell frequency.

Jackknife Method. Fleiss and Davies (1982) derived the Jackknife estimator of Cohen's kappa coefficient, $\hat{\kappa}_J$, obtained by a weighted average of pseudo-values. The pseudo-values are defined as $\widetilde{\kappa}_{jk} = N\hat{\kappa} - (N-1)\hat{\kappa}_{-jk}$ where $\hat{\kappa}_{-jk}$ is Cohen's kappa coefficient obtained when one unit is deleted from the cell (j,k), $j,k = 1,\cdots,K$. The Jackknife estimator of Cohen's kappa coefficient is then

$$\hat{\kappa}_J = \frac{1}{N}\sum_{j=1}^{K}\sum_{k=1}^{K} n_{jk}\widetilde{\kappa}_{jk} \tag{2.26}$$

and the estimated variance is

$$var(\hat{\kappa}_J) = \frac{1}{N(N-1)} \sum_{j=1}^{K} \sum_{k=1}^{K} n_{jk}(\widetilde{\kappa}_{jk} - \hat{\kappa}_J)^2. \tag{2.27}$$

The Jackknife procedure is described in general in Appendix B.

Bootstrap Method. The large sample variance of the Cohen's kappa coefficient can be determined by the bootstrap method as explained in Appendix B, by taking the variance of the bootstrapped coefficients.

2.3.5 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 (HIV). To assess the importance of cervical ectopy, methods for measuring ectopy with precision are needed. A computerized planimetry method was developed for measuring cervical ectopy and the reliability of that method was compared with direct visual assessment in a study conducted by Gilmour et al. (1997). Photographs of the cervix of 85 women without cervical disease were assessed for cervical ectopy by three medical raters who used both assessment methods. The response of interest, cervical ectopy size, was an ordinal variable with four categories: (1) minimal, (2) moderate, (3) large and (4) excessive. The classification of the 85 women by 2 of the 3 raters is summarized in Table 2.7 for the direct visual assessment in terms of frequency.

Table 2.7. 4×4 contingency table resulting from the direct visual assessment of cervical ectopy size by 2 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

Overall, the observed proportion of agreement is equal to

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

The proportion of agreement expected by chance is equal to

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

The two medical raters agree on 50.6% of the patients and agreement due to chance amounts 24.7%. This leads to a Cohen's kappa coefficient of

$$\hat{\kappa} = (0.506 - 0.247)/(1 - 0.247) = 0.343.$$

The maximum observed proportion of agreement permitted by the marginals is equal to $p_{oM} = (15 + 29 + 13 + 11)/85 = 0.800$, leading a maximum value of Cohen's kappa coefficient of $\hat{\kappa}_M = 0.734$.

To determine the agreement on each category, 2×2 tables were constructed by isolating one category and collapsing all the other categories together. They are represented in Table 2.8. The corresponding observed proportions of agreement (p_o), proportions of agreement expected by chance (p_e), Cohen's kappa coefficients ($\hat{\kappa}$), maximum observed proportion of agreement permitted by the marginal (p_{oM}) and the resulting kappa coefficient $\hat{\kappa}_M$ are also provided.

Table 2.8. 2×2 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 4-categorical scale

Category Minimal			
	Rater 2		
Rater 1	Minimal	Other	Total
Minimal	13	2	15
Other	14	56	70
Total	27	58	85

Category Moderate			
	Rater 2		
Rater 1	Moderate	Other	Total
Moderate	16	13	29
Other	13	43	56
Total	29	56	85

Category Large			
	Rater 2		
Rater 1	Large	Other	Total
Large	3	10	13
Other	15	57	72
Total	18	67	85

Category Excessive			
	Rater 2		
Rater 1	Excessive	Other	Total
Excessive	11	17	28
Other	0	57	57
Total	11	74	85

As seen in Table 2.9, the agreement on extreme categories (Minimal and Excessive) is better than agreement on middle categories (Moderate and Large). This is a well-know phenomenon. It is easier to distinguish between extreme categories than middle ones. The agreement on category Large is almost nil while the agreement on category Excessive is the maximal agreement permitted by the marginals.

Table 2.9. Observed proportions of agreement (p_o), proportions of agreement expected by chance (p_e), Cohen's kappa coefficients ($\hat{\kappa}$), maximum observed proportions of agreement permitted by the marginal (p_{oM}) and the resulting kappa coefficients $\hat{\kappa}_M$ relative to the tables given in Table 2.8

Category	p_o	p_e	$\hat{\kappa}$	p_{oM}	$\hat{\kappa}_M$
Minimal	0.812	0.618	0.507	0.859	0.631
Moderate	0.694	0.550	0.320	1.0	1.0
Large	0.706	0.700	0.019	0.941	0.803
Excessive	0.800	0.626	0.465	0.800	0.465
Overall	0.506	0.247	0.343	0.800	0.734

2.4 Intraclass kappa coefficient

2.4.1 Definition

Kraemer (1979) proposed to define kappa in terms of population parameters, by analogy to the intraclass correlation coefficient for continuous data, but adapted to the categorical case. The intraclass kappa coefficient can be viewed as a special case of Cohen's kappa coefficient where it is assumed that the ratings are interchangeable. In other words, the two raters are assumed to have the same marginal probability distribution. The resulting index is algebraically equivalent to Scott's index of agreement in the 2×2 case (Scott, 1955).

Consider again a population of items $\mathcal{I}$. Let $Y_{ij,r}$ be a random variable such that $Y_{ij,r} = 1$ if a randomly selected item i of population $\mathcal{I}$ is classified in category j ($j = 1, \cdots, K$) by rater r ($r = 1, 2$). Let $E(Y_{ij,r}) = \pi_j$, $(\pi'_j = 1 - \pi_j)$, expectations being taken over the population of items. The *intraclass kappa coefficient* relative to category j is defined by

$$\kappa_{I[j]} = \frac{cov(Y_{ij,1}, Y_{ij,2})}{\pi_j(1 - \pi_j)} \tag{2.28}$$

and the *global intraclass kappa coefficient* (κ_I) over the K categories is given by

$$\kappa_I = \frac{\sum_{j=1}^{K} cov(Y_{ij,1}, Y_{ij,2})}{\sum_{j=1}^{K} \pi_j(1 - \pi_j)}. \tag{2.29}$$

The intraclass kappa coefficient has the same form as Cohen's kappa coefficient,

$$\kappa_I = \frac{\Pi_{oI} - \Pi_{eI}}{1 - \Pi_{eI}} \tag{2.30}$$

where $\Pi_{oI} = \sum_{j=1}^{K} E(Y_{ij,1}Y_{ij,2}) = \sum_{j=1}^{K}(cov(Y_{ij,1}, Y_{ij,2}) + \pi_j^2)$

and $\Pi_{eI} = \sum_{j=1}^{K}[E(Y_{ij,1})E(Y_{ij,2})] = \sum_{j=1}^{K}\pi_j^2$.

2.4.2 Estimation of the parameters

Using the notations introduced in Table 2.1, the estimation of the intraclass kappa coefficient is obtained by replacing in the expression of κ_I Π_{oI} by $\hat{\Pi}_{oI} = p_{oI}$ with

$$p_{oI} = \sum_{j=1}^{K} p_{jj} \tag{2.31}$$

and Π_{eI} by $\hat{\Pi}_{eI} = p_{eI}$ with

$$p_{eI} = \sum_{j=1}^{K}\left(\frac{p_{j.} + p_{.j}}{2}\right)^2. \tag{2.32}$$

2.4.3 Properties for binary scales

Property 7. *Population model and maximum likelihood estimator*

As before (see Section 2.3.3, Property 6), consider the binary random variable $Y_{i,r}$. For item i, let $P(Y_{i,r} = 1) = E(Y_{i,r}|i) = P_i$ since the raters are assumed to be interchangeable. Over the population of items, let $E(P_i) = \pi$ and $var(P_i) = \sigma^2$. Then, the *intraclass kappa coefficient* can be rewritten as

$$\kappa_I = \frac{cov(Y_{i,1}, Y_{i,2})}{\sqrt{var(Y_{i,1})var(Y_{i,2})}} = \frac{E(Y_{i,1}Y_{i,2}) - \pi^2}{\pi(1-\pi)}, \tag{2.33}$$

if the data are summarized in a 2×2 contingency table (Table 2.10).

Table 2.10. Theoretical model for binary ratings made by 2 raters with equal marginal distributions

	Rater 2		
Rater 1	0	1	Total
0	$E[(1-Y_{i,1})(1-Y_{i,2})]$	$E[(1-Y_{i,1})Y_{i,2}]$	$1-\pi$
1	$E[Y_{i,1}(1-Y_{i,2})]$	$E[Y_{i,1}Y_{i,2}]$	π
Total	$1-\pi$	π	1

Using the expression of κ_I given by Equation 2.33, Table 2.10 can be rewritten (see Table 2.11). When the two discordant cells are grouped together, this table

expresses the probabilities of a model, known as the *common correlation model* (Bloch and Kraemer, 1989).

Table 2.11. Expected probability of joint responses for the classification of two raters on a binary scale (common correlation model)

	Rater 2		
Rater 1	0	1	Total
0	$(1-\pi)^2+\kappa_I\pi(1-\pi)$	$\pi(1-\pi)(1-\kappa_I)$	$1-\pi$
1	$\pi(1-\pi)(1-\kappa_I)$	$\pi^2+\kappa_I\pi(1-\pi)$	π
Total	$1-\pi$	π	1

Suppose the two raters classify a random sample of N items, leading to the contingency Table 2.2. The log-likelihood function is then

$$\begin{aligned}\ln L(\pi,\kappa_I|n_{11},n_{12},n_{21},n_{22}) &= n_{11}\ln[\pi^2+\kappa_I\pi(1-\pi)] \\ &+ (n_{12}+n_{21})\ln[\pi(1-\pi)(1-\kappa_I)] \\ &+ n_{22}\ln[(1-\pi)^2+\kappa_I\pi(1-\pi)]. \qquad (2.34)\end{aligned}$$

The maximum likelihood estimators of π and κ_I are

$$\hat{\pi}=\frac{2n_{11}+n_{12}+n_{22}}{2N}, \qquad (2.35)$$

$$\widehat{\kappa_I}=\frac{4(n_{11}n_{22}-n_{12}n_{21})-(n_{12}-n_{21})^2}{(2n_{11}+n_{12}+n_{21})(2n_{22}+n_{12}+n_{21})} \qquad (2.36)$$

which can be rewritten under the same form as Cohen's kappa coefficient, i.e., $\hat{\kappa}_I=(p_{oI}-p_{eI})/(1-p_{eI})$ with $p_{oI}=(n_{11}+n_{22})/N$ and $p_{eI}=\hat{\pi}^2+(1-\hat{\pi})^2$.

Property 8. *Effect of prevalence, sensitivity and specificity*

Kraemer (1979) showed the influence of the prevalence on the intraclass kappa coefficient in the binary case. Let M be a disease with prevalence noted $P=P(M)$ or equivalently the population of diseased subjects. Denote by $\overline{M}$ the population of non diseased subjects. Suppose that a test T is available such that a given subject will be declared "diseased" if the test is positive (T_+) and "non-diseased" if the test is negative (T_-). Let Y_i be random a variable such that $Y_i=1$ if the test is positive and $Y_i=0$ otherwise.

The *sensitivity* of the test is defined as the probability for a given subject to be declared positive if he/she is diseased. We have $S_e=P(T_+|M)=E_{|M}Y_i$.

In the same way, the *specificity* is defined as the probability for a given subject to be declared negative if he/she is disease free, $S_p = P(T_-|\overline{M}) = E_{|\overline{M}}(1 - Y_i)$.

Since

$$\pi = P(T_+) = P(T_+|M)P(M) + P(T_+|\overline{M})P(\overline{M}) = PS_e + (1 - P)(1 - S_p),$$

the probability π is the expectation of Y_i. Indeed,

$$E(Y_i) = E_{|M}Y_iP(M) + E_{|\overline{M}}Y_iP(\overline{M}) = PS_e + (1 - P)(1 - S_p) = \pi.$$

The variance of Y_i is

$$var(Y_i) = P(1 - P)(S_e + S_p - 1)^2.$$

From Equation 2.28, we have

$$\kappa_I = \frac{P(1 - P)(S_e + S_p - 1)^2}{\pi(1 - \pi)}. \tag{2.37}$$

It results that $\kappa_I = 1$ if and only if $S_e = 1$ and $S_p = 1$, i.e., T is a perfect test (pathognomonic test). If $P = 0$ or $P = 1$, $\kappa_I = 0$. Except for these extreme values, κ_I presents a maximum (κ_{IM}) if $P = \dfrac{\sigma_{S_p}}{\sigma_{S_e} + \sigma_{S_p}}$ where $\sigma^2_{S_p} = S_p(1 - S_p)$ and $\sigma^2_{S_e} = S_e(1 - S_e)$. By replacing the corresponding values in the expression of κ_I, we find

$$\kappa_{IM} = [(S_eS_p)^{1/2} - [(1 - S_e)(1 - S_p)]^{1/2}]^2. \tag{2.38}$$

Property 9. *Relation between Cohen's kappa and intraclass kappa coefficients*

When $K = 2$, both Cohen's kappa ($\hat{\kappa}$) and intraclass kappa ($\hat{\kappa_I}$) coefficients can be written under the same form:

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e} \tag{2.39}$$

with $p_o = (n_{11} + n_{22})/N$ and $p_e = (n_{1.}n_{.1} + n_{2.}n_{.2})/N^2$ and

$$\hat{\kappa}_I = \frac{p_{oI} - p_{eI}}{1 - p_{eI}} \tag{2.40}$$

with

$$p_{oI} = (n_{11} + n_{22})/N,$$
$$p_{eI} = ((2n_{11} + n_{12} + n_{21})/2N)^2 + ((2n_{22} + n_{12} + n_{21})/2N)^2.$$

It results that Cohen's kappa and the intraclass kappa are asymptotically equivalent. Indeed, since

$$p_{eI} - p_e = \frac{1}{2N^2}(n_{12} - n_{21})^2 \overset{N\to\infty}{\to} 0 \tag{2.41}$$

we have

$$|\hat{\kappa} - \hat{\kappa_I}| = \left| \frac{p_o - p_e}{1 - p_e} - \frac{p_o - p_{eI}}{1 - p_{eI}} \right| = \left| \frac{(p_e - p_{eI})(p_o - 1)}{(1 - p_c)(1 - p_{eI})} \right| \overset{N \to \infty}{\to} 0. \tag{2.42}$$

Moreover, it can be noted that Cohen's kappa and the intraclass kappa coefficient are equivalent when $n_{12} = n_{21}$, i.e., there is no rater bias.

2.4.4 Sampling variability

Delta method. Using the Delta method, the large sample variance of Cohen's kappa coefficient, with the additional assumption of homogeneous margins, simplifies to

$$\begin{aligned} var(\hat{\kappa_I}) &= \frac{1}{N(1 - C_3)^2} \left\{ \sum_{j=1}^{2} p_{jj}[1 - 4\overline{p}_j(1 - \hat{\kappa}_I)] \right. \\ &+ \left. (1 - \hat{\kappa}_I)^2 \sum_{j=1}^{2} \sum_{k=1}^{2} p_{jk}(\overline{p}_j + \overline{p}_k)^2 - [\hat{\kappa}_I - C_3(1 - \hat{\kappa}_I)]^2 \right\} \end{aligned} \tag{2.43}$$

where $p_{jk} = n_{jk}/N,\ j, k = 1, 21,\ p_{i.} = n_{i.}/N,\ p_{.j} = n_{.j}/N,\ \overline{p}_j = (p_{j.} + p_{.j})/2$ and $C_3 = \overline{p}_1 + \overline{p}_2$.

Bloch and Kraemer method. The expression of the standard error of the intraclass kappa obtained by the Delta method being quite unpleasant, Bloch and Kraemer (1989) instead proposed to use the formula derived by Fisher (1958). This method is based on the Taylor series expansion and led to

$$var(\hat{\kappa}_I) = \frac{(1 - \hat{\kappa}_I)}{N} \left[(1 - \hat{\kappa}_I)(1 - 2\hat{\kappa}_I) + \frac{\hat{\kappa}_I(2 - \hat{\kappa}_I)}{2\hat{\pi}(1 - \hat{\pi})} \right]. \tag{2.44}$$

Jackknife and bootstrap methods. The standard error of the intraclass kappa coefficient can also be derived by the Jackknife and the bootstrap method, in the same way as for Cohen's kappa coefficient (see Section 2.3.4).

2.4.5 Example

Pursuing with the cervical ectopy data obtained on 85 women by two raters, we calculated the intraclass kappa coefficient for each category of the 4-category scale (see Table 2.12).

For example, when considering the category "Minimal" against all other categories, the proportion of observed agreement is equal to

$$p_{oI} = p_o = \frac{13 + 56}{85} = 0.812$$

Table 2.12. Observed proportions of agreement (p_o), proportions of agreement expected by chance (p_{eI}) and, intraclass kappa coefficients ($\hat{\kappa}_I$) relative to the tables in Table 2.8

Category	p_{oI}	p_{eI}	$\hat{\kappa}_I$
Minimal	0.812	0.628	0.494
Moderate	0.694	0.550	0.320
Large	0.706	0.702	0.014
Excessive	0.800	0.646	0.434
Overall	0.506	0.263	0.330

but the proportion of agreement expected by chance differs, namely

$$p_{eI} = ((2 \times 13 + 2 + 14)/(2 \times 85))^2 + ((2 \times 56 + 2 + 14)/(2 \times 85))^2 = 0.628.$$

This leads to an intraclass kappa coefficient of

$$\hat{\kappa}_I = \frac{p_{oI} - p_{eI}}{1 - p_{eI}} = 0.494.$$

Remark that, when the marginal distribution of the two raters are the same (see Category Moderate in Table 2.12), we effectively have $\hat{\kappa} = \hat{\kappa}_I$. The overall intraclass coefficient is equal to $\hat{\kappa}_I = (0.506 - 0.263)/(1 - 0.263) = 0.330$.

2.5 Weighted kappa coefficient

2.5.1 Definition

Often some disagreements between the two raters can be considered as more important than others. For example, disagreement on two distant categories should be considered more important than on neighbouring categories on an ordinal scale. For this reason, Cohen (1968) 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 (see Table 2.1). The weighted kappa coefficient is defined in terms of agreement weights

$$\hat{\kappa}_w = \frac{p_{ow} - p_{ew}}{1 - p_{ew}} \tag{2.45}$$

with $p_{ow} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{jk}$ and $p_{ew} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{j.} p_{.k}$ ($0 \leq w_{jk} \leq 1$ and $w_{jj} = 1$).

It can also be defined with disagreement weights,

$$\hat{\kappa}_w = 1 - \frac{q_{ow}}{q_{ew}} \tag{2.46}$$

with $q_{ow} = \sum_{j=1}^{K}\sum_{k=1}^{K} v_{jk}p_{jk}$ and $q_{ew} = \sum_{j=1}^{K}\sum_{k=1}^{K} v_{jk}p_{j.}p_{.k}$ ($0 \leq v_{jk} \leq 1$ and $v_{jj} = 0$).

Although weights can be arbitrarily defined, two weighting schemes are most commonly used. These are the "linear" weights introduced by Cicchetti and Allison (1971)

$$w_{jk} = 1 - \frac{|j-k|}{K-1} \tag{2.47}$$

and the quadratic weights introduced by Fleiss and Cohen (1973)

$$w_{jk} = 1 - \left(\frac{|j-k|}{K-1}\right)^2. \tag{2.48}$$

Note that the disagreement weights $v_{jk} = (j-k)^2$ are also commonly used (Ludbrook, 2002; Agresti, 1992) and 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.

2.5.2 Properties

Quadratic weighting scheme. Cohen (1968) showed that if the marginal distribution of the two raters are the same and if the weights of disagreement are defined as $v_{jk} = (j-k)^2$, the weighted kappa coefficient is equivalent to the Pearson's correlation coefficient. This is a generalization of what was found for binary scales ($\hat{\kappa} = \hat{\phi}$). Furthermore, Fleiss and Cohen (1973) showed that using these weights v_{jk}, the weighted kappa coefficient has the same interpretation as the intraclass correlation coefficient of reliability when systematic variability between raters is included as a component of total variation. Finally, Schuster (2004) explicitly decomposed the weighted kappa coefficient defined with the weights v_{jk} in terms of rater means, rater variances and rater covariance in the context of a two-way ANOVA setting.

Consider the following two-way analysis of variance model. Let rater r ($r = 1, 2$) assign item i ($i = 1, \cdots, N$) in category k ($k = 1, \cdots, K$) and $Y_{i,r}$ denote the category score of item i.

$$Y_{i,r} = \mu + B_i + A_r + E_{i,r} \tag{2.49}$$

where B_i represents the random item effect, A_r the rater effect, either considered as fixed or random and $E_{i,r}$ the error term. Using the disagreement weights $v_{jk} = (j-k)^2$, Fleiss and Cohen (1973) shown that the weighted kappa coefficient can be rewritten as

$$\hat{\kappa}_w = \frac{BMS - EMS}{BMS + EMS + \frac{2}{N-1}JMS} \tag{2.50}$$

where BMS, JMS and EMS refer to item, rater and error mean squares, respectively based on $N-1$, $R-1$ and $(R-1)(N-1)$ degrees of freedom. Schuster (2004) showed that under the assumption of equal rater means, the weighted kappa coefficient is equivalent to ICC_{C2} or ICC_{C3} depending if raters are considered as random or fixed, respectively.

$$\hat{\kappa}_w = \frac{BMS - EMS}{BMS + (R-1)EMS} = \frac{BMS - EMS}{BMS + EMS}. \tag{2.51}$$

By additionally assuming equality of rater variances, Cohen (1968) showed that the weighted kappa coefficient is equivalent to Pearson's correlation coefficient. Schuster (2004) remarked that the Pearson's correlation coefficient thus represents an upper limit of the weighted kappa coefficient.

Linear weighting scheme. Vanbelle and Albert (2009c) revisited the weighted kappa coefficient with linear weights for ordinal scales to provide an intuitive interpretation of it. For any "cut-off" value k $(k = 1, \cdots, K-1)$, they reduced the $K \times K$ contingency table (see Table 2.1) into a 2×2 classification table by summing up all observations below and above the first k rows and first k columns (see Table 2.13) where

$$\begin{array}{ll} N_{11}(k) = \sum_{i=1}^{K}\sum_{j=1}^{K} n_{ij} & N_{12}(k) = \sum_{i=1}^{K}\sum_{j=k+1}^{K} n_{ij} \\ N_{21}(k) = \sum_{i=k+1}^{K}\sum_{j=1}^{K} n_{ij} & N_{22}(k) = \sum_{i=k+1}^{K}\sum_{j=k+1}^{K} n_{ij} \end{array}$$

Let $F_{lm}(k) = \frac{1}{N}N_{lm}(k)$, $F_{l.} = \frac{1}{N}N_{l.}(k)$ and $F_{.m} = \frac{1}{N}N_{.m}(k)$ be the corresponding joint and marginal frequencies $(l, m = 1, 2; k = 1, \cdots, K-1)$. Finally, denote by

$$p_o(k) = F_{11}(k) + F_{22}(k) \tag{2.52}$$

and

$$p_e(k) = F_{1.}(k)F_{.1}(k) + F_{2.}(k)F_{.2}(k) \tag{2.53}$$

the observed and expected proportions of agreement corresponding to Table 2.13.

Table 2.13. Reduction of the $K \times K$ contingency table into a 2×2 classification table by selecting a cut-off level k $(k = 1, \cdots, K)$ on the ordinal scale

	Rater 2		
Rater 1	$\leq k$	$> k$	Total
$\leq k$	$N_{11}(k)$	$N_{12}(k)$	$N_{1.}(k)$
$> k$	$N_{21}(k)$	$N_{22}(k)$	$N_{2.}(k)$
Total	$N_{.1}(k)$	$N_{.2}(k)$	N

Now, consider the quantities

$$p_o^* = \frac{1}{K-1} \sum_{k=1}^{K-1} p_o(k) \tag{2.54}$$

and

$$p_e^* = \frac{1}{K-1} \sum_{k=1}^{K-1} p_e(k). \tag{2.55}$$

Vanbelle and Albert (2009c) showed that $p_o^* = p_{ow}$ and $p_e^* = p_{ew}$ where p_{ow} and p_{ew} are respectively the "linearly" weighted observed and expected agreement, as defined by Cicchetti and Allison (1971) (see proof in Section 2.8). Specifically, they showed that the observed and expected agreements are merely the mean values of the corresponding proportions of all 2×2 tables obtained by collapsing the first k categories and last $K - k$ categories $(k = 1, \cdots, K-1)$ of the original $K \times K$ classification table, giving an intuitive interpretation of the linearly weighted kappa coefficient.

2.5.3 Sampling variability

The Delta method gives

$$\begin{aligned} var(\hat{\kappa}_w) &= \frac{1}{N\left(1-p_{ew}\right)^4} \{ \sum_{j=1}^{K} \sum_{k=1}^{K} p_{jk} \left[w_{jk} \left(1-p_{ew}\right) - \left(\overline{w}_{j.} + \overline{w}_{.k}\right) \left(1-p_{ow}\right) \right]^2 \\ &- \left(p_{ow} p_{ew} - 2p_{ew} + p_{ow}\right)^2 \} \end{aligned} \tag{2.56}$$

where $\overline{w}_{.j} = \sum_{m=1}^{K} w_{mj} p_{m.}$ and $\overline{w}_{k.} = \sum_{s=1}^{K} w_{ks} p_{.s}$. The large sample variance can also be derived by the Jackknife and the bootstrap method.

2.5.4 Example

In the cervical ectopy example (Gilmour et al., 1997), women were classified on a 4-category Likert scale. Disagreements between category 1 (Minimal) and 4 (Ex-

cessive) may be considered more important than disagreements between category 1 (Minimal) and 2 (Moderate). The linear and quadratic weights corresponding to a 4-category scale are given in Table 2.14.

Table 2.14. Linear (left) and quadratic (right) weighting schemes for a 4-category scale

	Rater 2			
Rater 1	1	2	3	4
1	1.00	0.67	0.33	0.00
2	0.67	1.00	0.67	0.33
3	0.33	0.67	1.00	0.67
4	0.00	0.33	0.67	1.00

	Rater 2			
Rater 1	1	2	3	4
1	1.00	0.89	0.56	0.00
2	0.89	1.00	0.89	0.56
3	0.56	0.89	1.00	0.89
4	0.00	0.56	0.89	1.00

To determine the linearly weighted kappa coefficient, consider Table 2.15. The weighted observed agreement is the sum of the elements obtained by multiplying the columns w_{ij} and p_{ij}. The weighed expected agreement is the sum of the elements obtained by multiplying the columns w_{ij} and $p_{i.}p_{.j}$.

The linearly weighted kappa coefficient was found to be 0.520 with $p_{ow} = 0.800$ and $p_{ew} = 0.583$ while the quadratic weighted kappa coefficient was equal to 0.666 ($p_{ow} = 0.907$ and $p_{ew} = 0.722$).

Table 2.15. Elements to determine the linearly weighted kappa coefficient for the cervical ectopy size example

Rater 1	Rater 2	w_{ij}	p_{ij}	$p_{i.}p_{.j}$	Rater 1	Rater 2	w_{ij}	p_{ij}	$p_{i.}p_{.j}$
1	1	1.00	0.15	0.05	3	1	0.33	0.04	0.04
1	2	0.67	0.02	0.05	3	2	0.67	0.08	0.04
1	3	0.33	0.00	0.03	3	3	1.00	0.04	0.03
1	4	0.00	0.00	0.02	3	4	0.67	0.00	0.02
2	1	0.67	0.12	0.09	4	1	0.00	0.01	0.09
2	2	1.00	0.19	0.10	4	2	0.33	0.05	0.10
2	3	0.67	0.04	0.06	4	3	0.67	0.14	0.06
2	4	0.33	0.00	0.04	4	4	1.00	0.13	0.04

2.6 Examples

2.6.1 Agreement and association

A frequent mistake is to use a chi-square test to quantify agreement between raters. The example of Fermanian (1984) illustrates this confusion. Let two raters classify independently $N = 100$ patients in three diagnostic categories A, B and C, leading to the contingency table displayed in Table 2.16 (first line). Under the hypothesis of independence of the two ratings, the expected cell counts are determined (second line of Table 2.16).

$$T_{jk} = \frac{n_{j.} n_{.k}}{N} \quad j, k = 1, 2, 3. \tag{2.57}$$

Table 2.16. Fermanian's example (1984): observed and expected cell counts

	Rater 2			
Rater 1	A	B	C	Total
A	16[a]	0	24	40
	16[b]	8	16	
B	20	6	4	30
	12	6	12	
C	4	14	12	30
	12	6	12	
Total	40	20	40	100

[a] Observed cell count

[b] Expected cell count

Under the hypothesis of independence between the two raters, the statistic

$$\chi^2 = \sum_{j=1}^{3} \sum_{k=1}^{3} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \tag{2.58}$$

where O_{jk} is the observed cell count in the cell (j, k) and E_{jk} is the corresponding expected cell count, follows a chi-square distribution with $(K-1)(K-1)$ degrees of freedom.

For the example, $\chi^2_{obs} = 38.7$ with 4 degrees of freedom. Hence, there is a highly significant association between the two ratings. However, the observed proportion of agreement is equal to $p_o = 0.34$ and the proportion of agreement expected by chance to $p_e = 0.34$. Thus, Cohen's kappa coefficient is equal to

$$\hat{\kappa} = \frac{0.34 - 0.34}{1 - 0.34} = 0.$$

Despite the existence of a strong association between the two ratings, agreement between the raters is only to be expected by chance. This example shows that agreement and association are different things.

2.6.2 Blood clots detection

A study was conducted on 50 patients to measure the efficacy of two new methods with respect to a standard method in the detection of blood clots in the legs (unpublished data). Each patient was classified as having (1) or not having (0) blood clot(s) in the legs with respect to a reference method called "Standard" and 2 new methods "Method 1" and "Method 2". Age and gender were also recorded for each patient. The study aimed at comparing the agreement between the standard method and each of the new methods in order to make a choice between them. There were 23 (46.0%) women and 27 (54.0%) men involved in the study. Their mean age was 69.0 ± 16.1 years (range: 32-97 years). The classification of the patients according to the presence of blood clots is given in Table 2.17 for the entire population and in Table 2.18 according to gender.

Table 2.17. Blood clots detection (0=No, 1=Yes) in the legs of 50 patients with a standard method and two new methods

		Method 1			Method 2		
		0	1	Total	0	1	Total
Standard	0	18	11	29	26	3	29
	1	4	17	21	4	17	21
	Total	22	28	50	30	20	50

Table 2.18. Blood clots detection (0=No, 1=Yes) in the legs of 23 women and 27 men with a standard method and two new methods

			Method 1			Method 2		
Gender	Method		0	1	Total	0	1	Total
Women	Standard	0	5	6	11	10	1	11
		1	0	12	12	1	11	12
		Total	5	18	23	11	12	23
Men	Standard	0	13	5	18	16	2	18
		1	4	5	9	3	6	9
		Total	17	10	27	19	8	27

Cohen's kappas corresponding to Tables 2.17 and 2.18 are given in Table 2.19 with their standard error (SE) determined by the Delta and the Jackknife methods.

Table 2.19. Blood clots detection example: Cohen's kappa coefficients ($\hat{\kappa} \pm SE$) for all patients and according to patients' gender

	All ($N = 50$)	Men ($N = 27$)	Women ($N = 23$)
Delta SE			
Method 1 - Standard	0.41±0.12	0.27±0.19	0.47±0.16
Method 2 - Standard	0.71±0.10	0.57±0.17	0.83±0.12
Jackknife SE			
Method 1 - Standard	0.41±0.13	0.27±0.20	0.47±0.17
Method 2 - Standard	0.71±0.10	0.57±0.18	0.83±0.12

Method 2 clearly gives better agreement with the Standard method than Method 1 and should thus, at this stage of the study, be preferred to Method 1.

2.6.3 Cervical ectopy size

Partial data of the study of Gilmour et al. (1997) were presented in Section 2.3.5. The classification of the cervical ectopy size of 85 women by the 2 medical raters using direct visual assessment and the computerized planimetry method is given in Table 2.20.

Table 2.20. Assessment of the cervical ectopy size (1=Minimal, 2=Moderate, 3=Large and 4=Excessive) of 85 women by 2 raters with the visual assessment and the computerized planimetry methods

	Visual assessment						Computerized planimetry					
	Rater 2						Rater 2					
Rater 1	1	2	3	4	Total	Rater 1	1	2	3	4	Total
1	13	2	0	0	15	1	30	1	1	0	32
2	10	16	3	0	29	2	7	25	3	0	35
3	3	7	3	0	13	3	1	4	1	1	7
4	1	4	12	11	28	4	0	1	2	8	11
Total	27	29	18	11	85	Total	38	31	7	9	85

The weighted kappa coefficients with quadratic weights corresponding to these classifications are given in Table 2.21. The agreement between the two medical raters was slightly higher with the planimetry method than with the direct visual assessment.

Table 2.21. Cervical ectopy example: weighted kappa coefficients with quadratic weights ($\hat{\kappa}_w \pm SE$)

	Visual assessment	Planimetry method
Delta SE	0.67±0.061	0.82±0.051
Jackknife SE	0.67±0.062	0.82±0.053

2.7 Discussion

In this chapter, measures of agreement between two raters on a K-categorical scale were introduced: Cohen's kappa, intraclass kappa and weighted kappa coefficients. Cohen's kappa coefficient is mainly used to quantify agreement on nominal scales, the intraclass kappa coefficient on binary scales when no rater bias is assumed and the weighted kappa coefficient on ordinal scales. These coefficients all possess the same property of being equal to 1 when agreement is perfect and equal to 0 when agreement is due to chance and will therefore be said to belong to the *kappa-like* family. However, as mentioned in the review of Banerjee et al. (1993), this family does not represent the only issue in the measurement of agreement between two raters on a categorical scale. Indeed, when the binary scale can be viewed as a dichotomization of an underlying continuous variable that is unidimensional with normal distribution, the tetrachoric correlation coefficient (TCC) (Pearson, 1900) is preferred. This may be the case, for example, for radiological assessment of pneumoconisis (normal/abnormal), which is assessed from chest radiographies displaying a profusion of small irregular opacities (Banerjee et al., 1993). Note that the TCC quantifies agreement in a different context and estimate, albeit related, different quantities (Kramer, 1997). Bennett et al. (1954) also derived the S agreement coefficient, assuming a uniform marginal distribution for both raters.

Several criticisms on kappa coefficients were formulated in the literature. Firstly, Thompson and Walter (1988), Feinstein and Cicchetti (1990), Cicchetti and Feinstein (1990) and Byrt et al. (1993) pointed out that Cohen's kappa coefficient is dependent on the prevalence of the trait under study which indicates a serious limitation when comparing values of Cohen's kappa coefficient among studies with varying prevalence. The dependence studied by Thompson and Walter (1988) was relative to the prevalence of the true latent binary variable under study, keeping sensitivity and specificity fixed, while Feinstein and Cicchetti (1990) studied the dependence of Cohen's kappa coefficient on observed marginal prevalences, keeping the proportion of observed agreement fixed. Indeed, it may appear surprising to find a low agreement when diagonal cells in the 2×2 contingency table show substantially greater frequency than the off-diagonal cells. However, Bloch and Kraemer (1989) and Vach (2005) criticized the results of Thompson and Walter

(1988) by noting that the dependence occurred only if one can change the prevalence without changing sensitivity and specificity, which is generally not the case. Moreover, Vach (2005) pointed out that the dependence studied by Feinstein and Cicchetti (1990) is simply a consequence of the purpose of Cohen's kappa coefficient. This was also noted by Hoehler (2000), who remarked that rater bias, by definition, indicates disagreement. The latter author added that kappa should never be adjusted for bias and prevalence, as made by Banerjee et al. (1993) and Lantz and Nebenzahl (1996). An alternative should be the use of the intraclass kappa coefficient, which ignores the bias existing between the raters. However, Zwick (1988) proposed to study the bias that may arise between the raters and to only assume no rater bias when it is plausible.

The use of weighted kappa coefficients was also criticized. The weights are generally given a priori and defined arbitrarily. In practice, the linear (Cicchetti and Allison, 1971) and quadratic (Fleiss and Cohen, 1973) weighting schemes are the most widely used. Quadratic weights have received much attention in the literature because of their practical interpretation. For instance, Fleiss and Cohen (1973) and Schuster (2004) showed that using the 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. In addition, Schuster (2004) noted that the weighted kappa coefficient is sensitive to change in location and scale of the scores, the intraclass correlation coefficient only to changes in scale while the Pearson's correlation coefficient is not sensitive to any change in location or scale and thus stressed the searcher to use the right coefficient according to the question of interest. On an other hand, Vanbelle and Albert (2009c) focused on the linearly weighted kappa coefficient defined by Cicchetti and Allison (1971) and strove to give an intuitive interpretation of it. Graham and Jackson (1993) observed that the value of the weighted kappa coefficient can vary considerably according to the weighting scheme used and henceforth may lead to different conclusions but guidelines for the selection of weights are inexistent unlike in Brenner and Kliebsch (1996), who demonstrated, using simulations, that with linear weights, the weighted kappa coefficient is less sensitive to the number of categories and should therefore be preferred when the number of categories of the ordinal scale is large. Finally, Roberts and McNamee (1998, 2005) developed a symmetric matrix of kappa-type coefficients to assess the agreement on an ordinal scale, arguing that collapsing all aspects of agreement into a single measure (i.e., the weighted kappa coefficient) may be not sufficient when categories are defined qualitatively. The matrix elements measure how well different parts of the scale may be distinguished from each other and a weighted kappa coefficient can be derived from the diagonal elements of the matrix.

Another criticism about the kappa-like family agreement indexes is that, like correlation coefficients, the interpretation of kappa statistics is not clear except for 0 and 1 values. Landis and Koch (1977b) therefore constructed a classification to appreciate the strength of agreement. This classification is widely used but should be avoided since its construction is totally arbitrary. It is preferable to consider a confidence interval to appreciate the value of a kappa estimate, the lower bound being often the only of interest. Several methods were derived to estimate the sampling variability of agreement coefficients belonging in the kappa-like family. Fleiss and Cuzick (1979) wrote '*Many human endeavors have been cursed with repeated failures before final success is achieved. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. The derivation of a correct standard error for kappa is a third*'. This is still the case. Only the Delta, Kraemer, Jackknife and Garner's methods were presented in this chapter because Blackman and Koval (2000) conducted a simulation study to compare the confidence intervals obtained for the intraclass kappa coefficient in the binary case based on these methods. As conclusion, they provided a guidance in selecting a method in small samples, showed in Table 2.22.

Table 2.22. Guidance table for constructing confidence interval in small samples for the intraclass kappa coefficient in the binary case

$\hat{\kappa}$	Interpretation	Prevalence	Sample size	Method
$0.0 \leq \hat{\kappa} < 0.2$	Slight	$0.1 < \hat{P} < 0.9$	$N \geq 20$	Kraemer, Delta
$0.2 \leq \hat{\kappa} < 0.4$	Fair	$0.1 \leq \hat{P} \leq 0.9$	$N \geq 20$	Kraemer, Delta, Jackknife
$0.4 \leq \hat{\kappa} < 0.6$	Moderate	$0.2 \leq \hat{P} \leq 0.8$	$20 \leq N < 40$	Garner
		$\hat{P} \leq 0.2; \hat{P} \geq 0.8$	$N \geq 40$	Jackknife
$0.6 \leq \hat{\kappa} < 1.0$	Substantial	$0.1 < \hat{P} < 0.9$	$N \geq 20$	Garner

However, in the literature, rules relative to the minimal sample size and marginal totals to justify the asymptotic approximation of the Delta method or Kraemer's method are not clear. Bloch and Kraemer (1989) and Donner and Eliasziw (1992), however, noted that the confidence interval obtained with Kraemer method is only reasonable with 'large' sample size that were not found to be attained in most of the agreement studies. Bloch and Kraemer (1989) therefore proposed a variance stabilizing transformation of the intraclass kappa coefficient or the use of the Jackknife estimator of the variance while Donner and Eliasziw (1992) proposed a procedure based on the χ^2 goodness of fit statistic for binary scales and extended the procedure to multinomial scales later (Donner and Eliasziw, 1997). Donner and Eliasziw (1992) found satisfactory results for $N = 25$ as Bloch and Kraemer (1989). However, both methods perform poorly when agreement or prevalence is

extreme (near 0 or near 1). Note that the Jackknife estimation of the variance was also proposed by Fleiss and Davies (1982). Cantor (1996) provided sample-size determination for Cohen's kappa coefficient in the binary case when variance is estimated by the Delta method. Nam (2000) proposed an alternative procedure, the score method, to derive confidence interval for the intraclass kappa coefficient and shown that the method performed better than the method of Donner and Eliasziw (1992) for small sample sizes.

Despite the disadvantages and limitations of Cohen's kappa coefficient, this index is popular due to its simplicity and wide applicability. It should just be known, that kappa mixes two sources of disagreement among raters, disagreement due to bias among raters and disagreement that occur because raters evaluate the items differently (Mielke and Berry, 2008).

2.8 Proofs

Equivalence 1. *If*

$$p_o^* = \frac{1}{K-1}\sum_{k=1}^{K-1} p_o(k)$$

and

$$p_e^* = \frac{1}{K-1}\sum_{k=1}^{K-1} p_e(k)$$

where $p_o(k)$ and $p_e(k)$ are defined in Equation 2.52 and 2.53, then $p_o^* = p_{ow}$ and $p_e^* = p_{ew}$, where p_{ow} and p_{ew} are respectively the "linearly" weighted observed and expected agreement, as defined by Cicchetti and Allison (1971).

Proof. Indeed, since

$$\begin{aligned}
p_o^* &= \frac{1}{K-1}\sum_{k=1}^{K-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} + \sum_{i=k+1}^{K}\sum_{j=k+1}^{K} p_{ij}\right) \\
&= \frac{1}{K-1}\sum_{k=1}^{K-1}\left(\sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij} - \sum_{i=1}^{k}\sum_{j=k+1}^{K} p_{ij} - \sum_{i=k+1}^{K}\sum_{j=1}^{k} p_{ij}\right) \\
&= \sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij} - \frac{1}{K-1}\sum_{k=1}^{K-1}\left(\sum_{i=1}^{k}\sum_{j=k+1}^{K} p_{ij} + \sum_{i=k+1}^{K}\sum_{j=1}^{k} p_{ij}\right)
\end{aligned}$$

and

$$\begin{aligned} p_o &= \sum_{i=1}^{K}\sum_{j=1}^{K}\left(1-\frac{|i-j|}{K-1}\right)p_{ij} \\ &= \sum_{i=1}^{K}\sum_{j=1}^{K}p_{ij} - \frac{1}{K-1}\sum_{i=1}^{K}\sum_{j=1}^{K}|i-j|p_{ij} \\ &= \sum_{i=1}^{K}\sum_{j=1}^{K}p_{ij} - \frac{1}{K-1}\sum_{i=1}^{K}\sum_{j=1}^{i}(i-j)p_{ij} - \frac{1}{K-1}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}(j-i)p_{ij}, \end{aligned}$$

it suffices to prove that

$$\sum_{k=1}^{K-1}\left(\sum_{i=1}^{k}\sum_{j=k+1}^{K}p_{ij} + \sum_{i=k+1}^{K}\sum_{j=1}^{k}p_{ij}\right) = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}(j-i)p_{ij} + \sum_{i=1}^{K}\sum_{j=1}^{i}(i-j)p_{ij}. \quad (2.59)$$

We have successively,

$$\sum_{k=1}^{K-1}\left(\sum_{i=1}^{k}\sum_{j=k+1}^{K}p_{ij} + \sum_{i=k+1}^{K}\sum_{j=1}^{k}p_{ij}\right) = \sum_{k=1}^{K-1}\sum_{i=1}^{k}\sum_{j=k+1}^{K}p_{ij} + \sum_{k=1}^{K-1}\sum_{i=k+1}^{K}\sum_{j=1}^{k}p_{ij}$$

$$\begin{aligned} &= \sum_{i=1}^{1}\sum_{j=2}^{K}p_{ij} + \sum_{i=1}^{2}\sum_{j=3}^{K}p_{ij} + \cdots + \sum_{i=1}^{K-1}\sum_{j=K}^{K}p_{ij} \\ &+ \sum_{i=2}^{K}\sum_{j=1}^{1}p_{ij} + \sum_{i=3}^{K}\sum_{j=1}^{2}p_{ij} + \cdots + \sum_{i=K}^{K}\sum_{j=1}^{K-1}p_{ij} \end{aligned}$$

$$\begin{aligned} = &\sum_{i=1}^{1}\sum_{j=2}^{K}p_{ij} + \sum_{i=1}^{1}\sum_{j=3}^{K}p_{ij} + \sum_{i=2}^{2}\sum_{j=3}^{K}p_{ij} + \cdots + \sum_{i=1}^{1}\sum_{j=K}^{K}p_{ij} + \sum_{i=2}^{K-1}\sum_{j=K}^{K}p_{ij} \\ + &\sum_{i=K}^{K}\sum_{j=1}^{1}p_{ij} + \sum_{i=2}^{K-1}\sum_{j=1}^{1}p_{ij} + \sum_{i=K}^{K}\sum_{j=1}^{2}p_{ij} + \sum_{i=3}^{K-1}\sum_{j=1}^{2}p_{ij} + \cdots + \sum_{i=K}^{K}\sum_{j=1}^{K-1}p_{ij} \end{aligned}$$

$$\begin{aligned} &= \sum_{j=2}^{K}(j-1)p_{1j} + \sum_{i=2}^{2}\sum_{j=3}^{K}p_{ij} + \cdots + \sum_{i=2}^{K-1}\sum_{j=K}^{K}p_{ij} \\ &+ \sum_{j=1}^{K-1}(K-j)p_{Kj} + \sum_{i=2}^{K-1}\sum_{j=1}^{1}p_{ij} + \cdots + \sum_{i=3}^{K-1}\sum_{j=1}^{2}p_{ij} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=2}^{K}(j-1)p_{1j} + \sum_{j=3}^{K}(j-2)p_{2j} + \cdots + \sum_{j=K}^{K}(j-(K-1))p_{K-1,j} \\
&+ \sum_{j=1}^{K-1}(K-j)p_{Kj} + \sum_{j=1}^{K-2}(K-1-j)p_{K-1,j} + \cdots \\
&+ \sum_{j=1}^{K-(K-1)}(K-(K-1)-j)p_{K-(K-1),j} \\
&= \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}(j-i)p_{ij} + \sum_{i=1}^{K}\sum_{j=1}^{i}(i-j)p_{ij}. \qquad (2.60)
\end{aligned}
$$

Thus, $p_o^* = p_{ow}$. The proof for $p_e^* = p_{ew}$ proceeds similarly by replacing p_{ij} by $p_{i.}p_{.j}$ $(i, j = 1, \cdots, K)$. ■

CHAPTER 3

Agreement between several raters

3.1 Introduction

While it is easy to define the agreement between two raters on a categorical scale for a given item (they agree or they don't agree), this is not the case when agreement is searched between several raters ($R > 2$). Indeed, agreement on a given item between R raters may be defined by an arbitrary choice along a continuum ranging from agreement between a pair of raters to agreement among all raters, i.e. a concordant classification between g raters ($g = 2, \cdots, R$). The most restrictive definition is to ask that all R raters agree on the categorization of the item (De Moivre's definition of agreement) and the less restrictive one is the pairwise definition of agreement, assuming that an agreement occurs if and only if two raters categorize the item consistently (Hubert, 1977). The pairwise definition of agreement was used by Fleiss (1971) and Davies and Fleiss (1982). In between, Conger (1980) developed a general framework, permitting to choose the definition of agreement on the continuum going from 2 to R, the g-wise agreement indexes ($g = 2, \cdots, R$), including the De Moivre's (R-wise) and pairwise (2-wise) definitions of agreement. Recently, Mielke and Berry (2008) proposed a weighted kappa coefficient between R raters using the De Moivre's definition of agreement.

However, Light (1971) used another definition of agreement. Specifically, he identified a rater among the R raters as the gold standard or the reference and defined agreement as a consistent classification between the standard (or reference) and another rater. Conger (1980) showed that this coefficient of agreement is equiva-

lent to taking the average of the Cohen's kappa coefficients determined between all the $R(R-1)$ possible pairs of raters among the R raters.

Finally, a third approach consists in developing an agreement coefficient based on models analogous to the ANOVA models for quantitative variables (Landis and Koch, 1977c). More recently, Schuster and Smith (2005) proposed a dispersion-weighted kappa framework for multiple raters (not shown here) to determine the degree of agreement between many raters. The resulting framework includes the 2-wise agreement index (Conger, 1980) and the agreement index developed by Landis and Koch (1977c) as special cases.

3.2 Intraclass correlation coefficients

3.2.1 One-way random effects ANOVA model

3.2.1.1 Binary scale

Definition. Landis and Koch (1977c) considered the case of R raters classifying independently N items on a binary scale ($K = 2$) when the items are not always classified by all the raters. This corresponds to the one-way random effects ANOVA model (see Chapter 1, Section 1.3.1). In this section, the number of raters classifying each item will first be considered to be constant and equal to R. This does not mean that the same R raters all classify each item. Based on the one-way random effects ANOVA model (see Chapter 1, Equation 1.7) and similarly to the quantitative case, Landis and Koch (1977c) proposed to define the agreement as the ratio of the between items variability and the total variability (see Chapter 1, Equation 1.8)

$$\kappa_{ICC_1} = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2}. \tag{3.1}$$

Estimation of the parameters. Suppose that item i is classified on a binary scale by R raters. For each item, the R raters are not necessarily the same ($i = 1, \cdots, N$). Consider the random variable $Y_{i,r}$ equal to 1 if rater r ($r = 1, \cdots, R$) classifies item i ($i = 1, \cdots, N$) in category 1 and equal to 0 otherwise. Denote by $y_{i,r}$ the observed value of the random variable $Y_{i,r}$. Let $n_i = \sum_{r=1}^{R} y_{i,r}$ be the number of raters among the R raters classifying item i in category 1 and $p_i = n_i/R$ be the corresponding proportion ($i = 1, \cdots, N$). Finally, let $p = \sum_{i=1}^{N} n_i/NR$ denote the overall proportion of items classified in category 1.

The between sum of squares is estimated by

$$BSS = \sum_{i=1}^{N}\sum_{r=1}^{R}(p_i - p)^2 = \sum_{i=1}^{N}\sum_{r=1}^{R}(\frac{n_i}{R} - p)^2 = \sum_{i=1}^{N}\frac{(n_i - Rp)^2}{R} \tag{3.2}$$

with $N-1$ degrees of freedom and the within sum of squares by

$$WSS = \sum_{i=1}^{N}\sum_{r=1}^{R}(y_{i,r} - p)^2 = \sum_{i=1}^{N}\frac{n_i(R-n_i)}{R} \tag{3.3}$$

with $N(R-1)$ degrees of freedom. The mean squares BMS and WMS are then respectively defined by

$$BMS = \frac{1}{N}\sum_{i=1}^{N}\frac{(n_i - Rp)^2}{R} \tag{3.4}$$

and

$$WMS = \frac{1}{N(R-1)}\sum_{i=1}^{N}\frac{n_i(R-n_i)}{R}. \tag{3.5}$$

Note that $BMS = BSS/N$ and not $BSS/(N-1)$ as it should be the case. This approximation was made by Fleiss (1981) provided that $N \geq 20$. The agreement coefficient, by analogy to the quantitative case, is defined by

$$\hat{\kappa}_{ICC_1} = \frac{BMS - WMS}{BMS + (R-1)WMS}. \tag{3.6}$$

After some elementary algebraic manipulation, Equation 3.6 can be expressed as

$$\hat{\kappa}_{ICC_1} = 1 - \frac{\sum_{i=1}^{N} n_i(R-n_i)}{NR(R-1)p(1-p)} = 1 - \frac{\sum_{i=1}^{N} p_i(1-p_i)}{N(R-1)p(1-p)}. \tag{3.7}$$

This agreement coefficient possesses the following properties:

1. If $p_i = p$ $(i = 1, \cdots, N)$, with $p \neq 0$ and $p \neq 1$, there is no more discordance within items than between items. In that case, $\hat{\kappa}_{ICC_1}$ takes its minimum value, i.e. $-1/(R-1)$.
2. If each proportion p_i is equal to 0 or is equal to 1, the agreement on the items is perfect and $\hat{\kappa}_{ICC_1} = 1$.

When there are only two raters $(R = 2)$, the proposed agreement coefficient $\hat{\kappa}_{ICC_1}$ reduces to the intraclass kappa coefficient $\hat{\kappa}_I$ defined in Chapter 2, Section 2.4.1.

Sampling variability. Fleiss et al. (1979) showed that under the null hypothesis $H_0 : \kappa_{ICC_1} = 0$,

$$var(\hat{\kappa}_{ICC_1}) = \frac{2}{NR(R-1)}. \tag{3.8}$$

Remark that $var(\hat{\kappa}_{ICC_1})$ is independent of the overall proportion of items classified in category 1 (i.e., p). Since this is only asymptotic and valid to test for null agreement, it is recommended to use the Jackknife estimator of the sampling variability instead.

Unequal number of raters per item. When the number of raters classifying each item is not constant and is equal to R_i $(i = 1, \cdots, N)$, the between mean squares and the within mean squares are respectively defined by

$$BMS = \frac{1}{N} \sum_{i=1}^{N} \frac{(n_i - R_i p)^2}{R_i} \tag{3.9}$$

and

$$WMS = \frac{1}{N(\overline{R}-1)} \sum_{i=1}^{N} \frac{n_i(R_i - n_i)}{R_i} \tag{3.10}$$

where $\overline{R} = \sum_{i=1}^{N} R_i/N$. The agreement coefficient, by analogy to the quantitative case, is estimated by

$$\hat{\kappa}_{ICC_1} = \frac{BMS - WMS}{BMS + (R_0 - 1)WMS} \tag{3.11}$$

where

$$R_0 = \overline{R} - \frac{\sum_{i=1}^{N}(R_i - \overline{R})^2}{N(N-1)\overline{R}}.$$

Fleiss (1981) remarked that when N is "large", R_0 and $\overline{R}$ are similar. By replacing R_0 by $\overline{R}$ in Equation 3.11, the agreement coefficient is estimated by

$$\hat{\kappa}_{ICC_1} = \frac{BMS - WMS}{BMS + (\overline{R} - 1)WMS}. \tag{3.12}$$

Fleiss and Cuzick (1979) showed that under the null hypothesis $H_0 : \kappa_{ICC_1} = 0$,

$$var(\hat{\kappa}_{ICC_1}) = \frac{1}{N(\overline{R}-1)^2 \overline{R}_H} \left(2(\overline{R}_H - 1) + \frac{(\overline{R} - \overline{R}_H)(1 - 4p(1-p))}{\overline{R}p(1-p)} \right) \tag{3.13}$$

where

$$\overline{R}_H = \frac{N}{\sum_{i=1}^{N} \frac{1}{R_i}}$$

is the harmonic mean of the number of observations per item. However, since this is only asymptotic and valid to test for null agreement, it is recommended to use the Jackknife estimator of the sample variance instead.

3.2.1.2 Nominal scale

Definition. Suppose that the number of categories on which the items are classified is equal to $K \geq 2$. Denote by p_j the overall proportion of ratings in category j and by $\hat{\kappa}_{ICC_{[j]}}$ the value of the intraclass correlation coefficient obtained when category j is isolated from the other $K-1$ categories ($j = 1, \cdots, K$). Landis and Koch (1977c) proposed to take the weighted average

$$\hat{\overline{\kappa}} = \frac{\sum_{j=1}^{K} p_j(1-p_j)\hat{\kappa}_{[j]}}{\sum_{j=1}^{K} p_j(1-p_j)} \tag{3.14}$$

as an overall measure of inter-rater agreement. This expression simplifies to

$$\hat{\overline{\kappa}}_{ICC_1} = 1 - \frac{NR^2 - \sum_{i=1}^{N}\sum_{j=1}^{K} n_{ij}^2}{NR(R-1)\sum_{j=1}^{K} p_j(1-p_j)} \tag{3.15}$$

where n_{ij} is the number of raters classifying item i ($i = 1, \cdots, N$) in category j ($j = 1, \cdots, K$) ($\sum_{j=1}^{K} n_{ij} = R$). An algebraically equivalent version of Equation 3.15 was first presented by Fleiss (1971), who showed explicitly how the intraclass correlation coefficient represents a chance-corrected measure of agreement (see Section 3.6 for proof).

Sampling variability. Fleiss et al. (1979) showed that

$$var(\hat{\overline{\kappa}}) = \frac{2\left[\left(\sum_{j=1}^{K} p_j(1-p_j)\right)^2 - \sum_{j=1}^{K} p_j(1-p_j)(1-2p_j)\right]}{[\sum_{j=1}^{K} p_j(1-p_j)]^2 NR(R-1)}. \tag{3.16}$$

3.2.1.3 Example

Conger (1980) considered the following hypothetical example. Suppose that 4 raters ($R = 4$) have to assign 10 subjects ($N = 10$) in 3 categories ($K = 3$). The data are presented in Table 3.1 and summarized in Table 3.2. Suppose that the 4 raters are not necessarily the same for all subjects.

When interest is on category 1, the between sum of squares is equal to

$$BSS = \frac{1}{4}\left\{(4 - 4\times 0.375)^2 + (2 - 4\times 0.375)^2 + \cdots + (0 - 4\times 0.375)^2\right\} = 4.125$$

Table 3.1. Category assignment of 10 subjects by 4 raters in 3 categories and number of subjects assigned in each category by each rater

	Subject										Category		
Rater	1	2	3	4	5	6	7	8	9	10	1	2	3
1	1	1	1	1	1	2	2	2	3	3	5	3	2
2	1	1	1	1	2	1	2	3	3	3	5	2	3
3	1	2	2	3	1	1	2	2	2	3	3	5	2
4	3	3	3	3	1	1	2	2	2	3	2	3	5

Table 3.2. Number of raters classifying each of 10 subjects into one of 3 categories

	Subject										
Category	1	2	3	4	5	6	7	8	9	10	p_j
1	3	2	2	2	3	3	0	0	0	0	0.375
2	0	1	1	0	1	1	4	3	2	0	0.325
3	1	1	1	2	0	0	0	1	2	4	0.300

and the within sum of squares is equal to

$$WSS = \frac{1}{4}\{3(4-3) + 2(4-2) + \cdots + 0(4-0)\} = 5.25.$$

This leads to $BMS = 4.125/10 = 0.413$ and $WMS = 5.25/10(4-1) = 0.175$, whence

$$\hat{\kappa}_{ICC_1} = \frac{0.413 - 0.175}{0.413 + (4-1)0.175} = 0.253.$$

In the same way, we obtained $\hat{\kappa}_{ICC_1} = 0.278$ and $\hat{\kappa}_{ICC_1} = 0.206$ for categories 2 and 3, respectively. Since

$$\sum_{j=1}^{3} p_j(1-p_j) = 0.375(1-0.375) + 0.325(1-0.325) + 0.300(1-0.300) = 0.664,$$

the overall agreement coefficient is equal to

$$\begin{aligned} \hat{\bar{\kappa}}_{ICC_1} &= \frac{0.375(1-0.375)0.253 + 0.325(1-0.325)0.278 + 0.300(1-0.300)0.206}{0.664} \\ &= 0.247. \end{aligned}$$

3.2.2 Two-way ANOVA models

3.2.2.1 Binary scale

Definition. Similarly to the approach of Landis and Koch (1977c), it is possible to determine an agreement index when each item is rated on a binary scale by the same group of R raters, considered as fixed or randomly selected from a larger population (see Chapter 1, Section 1.3.2). The agreement index takes into account the systematic source of variation between items and between raters and is defined by analogy to the quantitative case by

$$\kappa_{ICC2} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_A^2 + \sigma_I^2 + \sigma_E^2} \text{ and } \kappa_{ICC2} = \frac{\sigma_B^2 - \sigma_I^2/(R-1)}{\sigma_B^2 + \theta_A^2 + \sigma_I^2 + \sigma_E^2}. \tag{3.17}$$

in case of random raters and fixed raters, respectively.

Estimation of the parameters. The coefficient κ_{ICC2}, in case of random and fixed raters, is estimated by

$$\hat{\kappa}_{ICC_2} = \frac{BMS - EMS}{BMS + (R-1)EMS + R(JMS - EMS)/N}. \tag{3.18}$$

We have, by analogy to the quantitative case,

$$BMS = \frac{R}{N-1}\sum_{i=1}^{N}(p_i - p)^2 = \frac{R}{N-1}\sum_{i=1}^{N}(p_i^2 - p^2), \tag{3.19}$$

$$JMS = \frac{N}{R-1}\sum_{r=1}^{R}(\overline{y}_{.,r} - p)^2 = \frac{N}{R-1}\sum_{r=1}^{R}(\overline{y}_{.,r}^2 - p^2), \tag{3.20}$$

$$EMS = \frac{1}{(N-1)(R-1)}[\sum_{i=1}^{N}\sum_{r=1}^{R}(y_{i,r} - \overline{y}_{.,r})^2 - R\sum_{i=1}^{N}(p_i - p)^2]. \tag{3.21}$$

where p_i and p were defined previously and $\overline{y}_{.,r} = \sum_{i=1}^{N} y_{i,r}/N$. After some elementary algebraic manipulation, Equation 3.18 can be expressed as

$$\begin{aligned} \hat{\kappa}_{ICC_2} &= 1 \\ &- \frac{(N-1)\sum_{i=1}^{N} n_i(R-n_i)}{R^2\sum_{i=1}^{N}(p_i-p)^2 - N\sum_{i=1}^{N}\sum_{r=1}^{R}(y_{i,r}-\overline{y}_{.,r})^2 + NR^2(N-1)p(1-p)}. \end{aligned} \tag{3.22}$$

Sampling variability. It is suggested to use the Jackknife estimator to determine the sampling variability.

3.2.2.2 Nominal scale

Definition. The approach of Landis and Koch (1977c) defined in the one-way ANOVA setting can be applied to the two-way ANOVA setting (see Equation 3.14), leading

$$\hat{\bar{\kappa}}_{ICC_2} = \frac{\sum_{j=1}^{K} p_j(1-p_j)\hat{\kappa}_{ICC_{2[j]}}}{\sum_{j=1}^{K} p_j(1-p_j)}. \tag{3.23}$$

Sampling variability. It is suggested to determine the sampling variability using the Jackknife technique.

3.2.2.3 Example

Suppose now that in the hypothetical example of Conger (1980), each subject is classified by the same set of 4 raters in the 3 categories. To calculate the overall agreement index over the 3 categories, the agreement indexes are needed for each category. The ANOVA table relative to category 1 is given in Table 3.3.

Table 3.3. Two-way ANOVA table when interest is on category 1 for the example of Conger (1980)

Variability	Sum of squares	Degrees of freedom	Mean squares
Between items	4.125	9	0.458
Within items	131.083	30	4.369
Between raters	129.708	3	0.169
Residuals	9.275	27	0.140
Total	135.208	39	

The intraclass correlation coefficient relative to category 1 is thus equal to

$$\hat{\kappa}_{ICC_2} = \frac{0.458 - 0.169}{0.458 + (4-1)0.169 + 4(0.492 - 0.169)/10} = 0.292.$$

In the same way, the intraclass coefficient relative to categories 2 and 3 are equal to 0.302 and 0.280, respectively, leading an overall agreement index of

$$\begin{aligned} \hat{\kappa}_{ICC_2} &= \frac{0.375(1-0.375)0.313 + 0.325(1-0.325)0.302 + 0.300(1-0.300)0.280}{0.664} \\ &= 0.334. \end{aligned}$$

3.3 g-wise agreement indexes

3.3.1 General framework

Definition. Suppose that R raters have to classify N items on a K-categorical scale and that agreement is defined as a consistent classification of g raters ($g \leq R$). Let n_{ij} denote the number of raters classifying item i in category j ($i = 1, \cdots, N; j = 1, \cdots, K$), $p_{j,r}$ the proportion of items assigned in category j by rater r and $p_j = \sum_{r=1}^{R} p_{j,r}/R$ the overall proportion of items classified in category j ($j = 1, \cdots, K$). The g-wise observed proportion of agreement is defined as

$$p_o(g) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{K} \prod_{r=0}^{g-1} (n_{ij} - r)}{N \prod_{r=0}^{g-1} (R - r)} \tag{3.24}$$

and the proportion of g-wise agreement expected by chance is equal to

$$p_e(g) = \frac{1}{C_R^g} \sum_{r^{(1)} < \cdots < r^{(g)}} \sum_{j=1}^{K} \prod_{s=1}^{g} \overline{y}_{j,r^{(s)}} \tag{3.25}$$

where $\sum_{r^{(1)} < \cdots < r^{(g)}}$ is the summation over all g-tuples of raters such that $1 \leq r^{(1)} < \cdots < r^{(g)} \leq R$. This leads to the g-wise agreement index

$$\hat{\kappa}(g) = \frac{p_o(g) - p_e(g)}{1 - p_e(g)}. \tag{3.26}$$

The intraclass version is obtained by considering

$$p_e(g) = \sum_{j=1}^{K} p_j^g. \tag{3.27}$$

Sampling variability. It is suggested to use the Jackknife estimator of the standard error for the g-wise agreement indexes.

3.3.2 Pairwise agreement index

Definition. Davies and Fleiss (1982) proposed a chance-corrected measure of agreement based on pairwise agreement which can be expressed in terms of mean squares under a two-way ANOVA setting when the scale is binary and is equivalent to the 2-wise agreement index introduced by Conger (1980). Suppose that each of several raters ($r = 1, \cdots, R$) classify each of a sample of items ($i = 1, \cdots, N$) on a K-categorical scale. Let the random variable $Y_{ij,r}$ equal to 1 when rater r classifies item i in category j ($\sum_{j=1}^{K} Y_{ij,r} = 1$) and $y_{ij,r}$ denote the achievement of the random variable $Y_{ij,r}$. Finally, let $n_{ij} = \sum_{r=1}^{R} y_{ij,r}$ be the number of raters classifying item i in category j ($i = 1, \cdots, N; j = 1, \cdots, R$). Davies and Fleiss (1982) defined the

observed proportion of agreement as the mean observed proportion of agreement between all $R(R-1)$ possible pairs of raters among the R raters.

$$\begin{aligned} p_o &= \frac{1}{R(R-1)} \sum_{r=1}^{R} \sum_{r' \neq r} o_{r,r'} \\ &= \frac{1}{NR(R-1)} \sum_{i=1}^{N} \sum_{j=1}^{K} \sum_{r=1}^{R} \sum_{r' \neq r} y_{ij,r} y_{ij,r'} \\ &= \frac{1}{NR(R-1)} \sum_{i=1}^{N} \sum_{j=1}^{K} n_{ij}(n_{ij}-1) \\ &= \frac{1}{NR(R-1)} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{K} n_{ij}^2 - NR \right\}. \end{aligned} \tag{3.28}$$

In the same way,

$$\begin{aligned} p_e &= \frac{1}{R(R-1)} \sum_{r=1}^{R} \sum_{r' \neq r} e_{r,r'} \\ &= \frac{1}{R(R-1)} \sum_{j=1}^{K} \sum_{r=1}^{R} \sum_{r' \neq r} \overline{y}_{j,r} \overline{y}_{j,r'} \\ &= \sum_{j=1}^{K} p_j^2 - \frac{1}{R(R-1)} \sum_{j=1}^{K} \sum_{r=1}^{R} (\overline{y}_{j,r} - p_j)^2 \end{aligned} \tag{3.29}$$

where $p_j = \sum_{r=1}^{R} \overline{y}_{j,r}/R$ is the overall proportion of items classified in category j $(j = 1, \cdots, K)$. Davies and Fleiss (1982) then defined the agreement coefficient

$$\begin{aligned} \hat{\kappa}_D &= \frac{p_o - p_e}{1 - p_e} \\ &= 1 - \frac{NR^2 - \sum_{i=1}^{N} \sum_{j=1}^{K} n_{ij}^2}{N\{R(R-1) \sum_{j=1}^{K} p_j(1-p_j) + \sum_{j=1}^{K} \sum_{r=1}^{R} (\overline{y}_{j,r} - p_j)^2\}}. \end{aligned} \tag{3.30}$$

When there are only two raters ($R = 2$), the agreement coefficient defined by Equation 3.30 reduces to Cohen's kappa coefficient (see Chapter 2, Section 2.3). The pairwise agreement index proposed by Davies and Fleiss (1982) is equivalent to the 2-wise agreement index of Conger (1980).

When interest lies in only one category j, all the other categories than the category of interest may be combined into a single category and the problem reduces to the binary case. The resulting agreement coefficient proposed by Davies and Fleiss (1982) then simplifies to

$$\hat{\kappa}_{D[j]} = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{\sum_{i=1}^{N} n_{ij}(R - n_{ij})}{N\{R(R-1)p_j(1-p_j) + \sum_{r=1}^{R} (\overline{y}_{j,r} - p_j)^2\}} \tag{3.31}$$

and can be expressed in terms of mean squares

$$\hat{\kappa}_{D[j]} = \frac{BMS - EMS}{BMS + (R-1)EMS + R\ JMS/(N-1)} \tag{3.32}$$

where BMS, EMS and JMS were defined previously.

However, the agreement coefficient relative to the two-way analysis of variance is given by Equation 3.18. Davies and Fleiss (1982) observed the equivalence of $\hat{\kappa}_{D[j]}$ and $\hat{\kappa}_{ICC_2}$ provided that N is large ($N > 15$).

Sampling variability. Davies and Fleiss (1982) only gave the formula of the standard error for the binary case and proposed a FORTRAN program for the nominal case since the form is to complicated. The Jackknife estimator to compute the sampling variance of this agreement index may therefore be an interesting alternative.

3.3.3 Weighted R-wise agreement index

Definition. Recently, Mielke and Berry (2008) introduced a weighted agreement index between R raters using the De Moivre's definition of agreement. Although their method is valid for any number R of raters, the method is presented for three raters ($R = 3$) for notation convenience. Let n_{jkl} be the number of items classified in category j by rater 1, category k by rater 2 and category l by rater 3 and $p_{j,r}$ the proportion of items assigned in category j by rater r ($j = 1, \cdots, K; r = 1, 2, 3$). The weighted observed agreement is defined by

$$p_{ow}(3) = \frac{1}{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{l=1}^{K} w_{jkl} n_{jkl} \tag{3.33}$$

and the weighted agreement expected by chance is defined by

$$p_{ew}(3) = \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{l=1}^{K} w_{jkl} p_{j,1} p_{k,2} p_{l,3} \tag{3.34}$$

leading to the weighted agreement index

$$\kappa_w(3) = \frac{p_{ow}(3) - p_{ew}(3)}{1 - p_{ew}(3)}. \tag{3.35}$$

Mielke and Berry (2008) proposed weighting schemes corresponding to the linear and the quadratic weighting schemes introduced in the case of two single raters. In case of three raters, the linear weighting scheme writes

$$w_{jkl} = |j - k| + |j - l| + |k - l| \ (j, k, l = 1, \cdots, K) \tag{3.36}$$

and the quadratic weighting scheme is

$$w_{jkl} = (j-k)^2 + (j-l)^2 + (k-l)^2 \ (j,k,l = 1, \cdots, K). \tag{3.37}$$

When the weights are defined to be

$$\begin{cases} w_{jkl} = 1 \text{ if } j = k = l \ (j,k,l = 1,2,3) \\ w_{jkl} = 0 \text{ otherwise,} \end{cases}$$

the weighted agreement index of Mielke and Berry (2008) is equivalent to the 3-wise agreement index of Conger (1980). This is also the case for $R > 3$.

Sampling variability. Mielke and Berry (2008) proposed to use an exact permutation test to test hypotheses. The procedure consists in generating all possible arrangements of the N items in the K^R cells of the contingency table resulting from the classification of the R raters, while preserving the marginal totals. For each arrangement of the cell frequencies, the weighted agreement index is determined. The number of times that the resulting weighted agreement indexes exceed or are equal to the value of the weighted agreement obtained from the original sample is then recorded. If this number, divided by the total number of permutations, is less than or equal to the α confidence level, then the null hypothesis is rejected. Since the number of permutations is usually very large for multi-way contingency tables, Mielke and Berry (2008) proposed to calculate the weighted agreement index for a large number (e.g., 1 000 000) of random tables.

3.3.4 Example.

Consider again the example of Conger (1980). To calculate the 2-wise (pairwise) agreement index, we need

$$p_o = \frac{1}{10 \times 4(4-1)}(100 - 10 \times 4) = 0.500$$

and

$$\begin{aligned} p_e &= 0.375^2 + 0.325^2 + 0.300^2 \\ &- \frac{\{(0.500-0.375)^2 + \cdots + (0.500-0.300)^2\}}{4(4-1)} = 0.322. \end{aligned}$$

This leads to

$$\hat{\kappa}_D = \frac{0.500 - 0.322}{1 - 0.322} = 0.263.$$

Then, to determine the 3-wise agreement index, we need

$$\begin{aligned} p_o(3) &= \frac{1}{10(4-0)(4-1)(4-2)} \\ &\times [(3-0)(3-1)(3-2)+(0-0)(0-1)(0-2)+(1-0)(1-1)(1-2) \\ &+ \cdots + (0-0)(0-1)(0-2)+(0-0)(0-1)(0-2) \\ &+ (4-0)(4-1)(4-2)] \\ &= 0.300 \end{aligned}$$

and

$$\begin{aligned} p_e(3) &= \frac{1}{4}[(0.5\times 0.5\times 0.3+0.3\times 0.2\times 0.5+0.2\times 0.3\times 0.2)+\cdots \\ &+ (0.5\times 0.3\times 0.2+0.2\times 0.5\times 0.3+0.3\times 0.2\times 0.5)] = 0.100. \end{aligned}$$

This leads to

$$\hat{\kappa}(3) = \frac{0.300-0.100}{1-0.100} = 0.222.$$

In the same way, the 4-wise agreement index is equal to

$$\hat{\kappa}(4) = \frac{0.200-0.039}{1-0.039} = 0.175.$$

3.4 Syphilis serology

A proficiency testing program for syphilis serology was conducted by the College of American Pathologists (CAP). For the fluorescent treponemal antibody absorption test (FTA-ABS), 3 reference laboratories were identified and considered as experts in the use of that test. During 1974, 40 syphilis serology specimens were tested independently by the 3 reference laboratories. Williams (1976) presented results obtained by the 3 reference laboratories and an additional participant (noted L) for 28 specimens (see Appendix A.2, Table A.2). Each specimen was classified as non-reactive (NR), borderline (BL) or reactive (RE). The different agreement coefficients between the 3 reference laboratories are given in Table 3.4 with their standard error derived by the Jackknife technique.

All the agreement indexes gave similar results. There is agreement between the 3 reference laboratories. As expected, the 2-wise agreement index is equal to the agreement index derived by Davies and Fleiss (1982). Although, Light's and the 2-wise agreement indexes seemed to be equal, Light's agreement index is equal to 0.67932 while the 2-wise agreement index is equal to 0.67908. It should be noted that it is not correct to compute the agreement index $\hat{\kappa}_{ICC_1}$ since the same 3 reference laboratories classified all specimens.

Table 3.4. Agreement coefficients obtained for the classification of 28 specimen by 3 reference laboratories

Coefficient	Estimate	SE
$\hat{\kappa}_{ICC_1}$	0.676	0.099
$\hat{\kappa}_{ICC_2}$	0.684	0.096
$\hat{\kappa}(2)$	0.679	0.097
$\hat{\kappa}(3)$	0.697	0.095
$\hat{\kappa}_D$	0.679	0.097
$\hat{\kappa}_{Light}$	0.679	0.097

3.5 Discussion

Several approaches were presented in this chapter. Firstly, agreement was defined as the ratio of the between items variability and the total variability by Landis and Koch (1977c). These agreement indexes are all based on the absolute definition and not on the consistency definition of agreement. The model on which the agreement coefficient is constructed should be chosen with care because the results have to be interpreted differently, depending on the model, as it was the case on quantitative scales. To our knowledge, only the coefficients derived by Landis and Koch (1977c) and Davies and Fleiss (1982) are usually used in practice although agreement indexes might be constructed on other ANOVA models.

As an alternative, agreement was defined as a concordant classification of g raters among the R raters ($g \leq R$). Conger (1980) defined a general framework, the g-wise agreement indexes, including the less restrictive (2-wise) and the most restrictive (R-wise) definition of agreement. No guideline was provided to determine the optimal number of raters g, which depends on the total number of raters R. The larger the number of raters, the more difficult it will be to have a concordant classification between all the raters. In practice, g is often determined to correspond to the majority of the raters ($g > R/2$). Note that Mielke and Berry (2008) preferred the R-wise rather than the 2-wise definition of agreement because all interactions between the R raters are not taken into account in the 2-wise definition.

Although the agreement coefficients were developed in different frameworks, the agreement coefficient proposed by Landis and Koch (1977c) in a one-way ANOVA framework is equivalent to the pairwise agreement coefficient derived by Fleiss (1971) and the agreement coefficient developed by Davies and Fleiss (1982) can be expressed in terms of mean squares in a two-way ANOVA setting. Note that this latter coefficient is also equivalent to the 2-wise agreement coefficient developed by

Conger (1980). In the present example, all approaches led to similar results. To our knowledge, guidelines for choosing between the different definitions of agreement do not exist and might therefore be the incentive for further research.

3.6 Proofs

Equivalence 2.

$$\hat{\bar{\kappa}}_{ICC_1} = 1 - \frac{NR^2 - \sum_{i=1}^{N}\sum_{j=1}^{K} n_{ij}^2}{NR(R-1)\sum_{j=1}^{K} p_j(1-p_j)} \tag{3.38}$$

can be expressed as a chance-corrected measure of agreement,

$$\hat{\bar{\kappa}}_{ICC_1} = \frac{p_o - p_e}{1 - p_e}. \tag{3.39}$$

Proof. Indeed, the number of pairs in agreement out of all $R(R-1)$ possible pairs is

$$p_{oi} = \frac{1}{R(R-1)}\sum_{j=1}^{K} n_{ij}(n_{ij}-1) = \frac{1}{R(R-1)}(\sum_{j=1}^{K} n_{ij}^2 - R). \tag{3.40}$$

Fleiss (1971) then defined the overall proportion of agreement as

$$p_o = \frac{1}{N}\sum_{i=1}^{N} p_{oi} = \frac{1}{NR(R-1)}(\sum_{i=1}^{N}\sum_{j=1}^{K} n_{ij}^2 - NR) \tag{3.41}$$

and the proportion of agreement expected by chance by

$$p_e = \sum_{j=1}^{K} p_j^2. \tag{3.42}$$

Simple algebraic manipulations show that

$$\hat{\kappa}_{ICC_1} = 1 - \frac{NR^2 - \sum_{i=1}^{N}\sum_{j=1}^{K} n_{ij}^2}{NR(R-1)\sum_{j=1}^{K} p_j(1-p_j)} = \frac{p_o - p_e}{1 - p_e}. \tag{3.43}$$

■

CHAPTER 4

Agreement between an isolated rater and a group or raters

4.1 Introduction

Cohen (1960) introduced the kappa coefficient $\hat{\kappa} = (p_o - p_e)/(1 - p_e)$ to quantify the agreement between two raters classifying items on a categorical scale. He corrected the proportion of items with concordant classification (p_o) for the proportion of concordant pairs expected by chance (p_e) and standardized the quantity to obtain 1 in case of perfect agreement between the two raters and 0 when the raters agree by chance. There are situations where agreement is searched between an isolated rater and a group of raters, regarded as a whole, a reference, expert or gold standard group, in which all raters may not perfectly agree with each other. For example, each of a series of candidates may be assessed against a group of experts with the purpose of evaluating their knowledge and classifying the candidates. This is a frequent exercise in education or in competence examinations. In the context of accreditation, a routine laboratory may have to reach a pre-defined level of agreement when challenged against a set of reference laboratories for a number of test specimens. Acknowledgment has to be make for the fact that the reference laboratories exhibit themselves analytical variability and do not necessarily agree with each other. The traditional approach to solve this problem is to determine a consensus in the group of raters and to measure the agreement between the isolated rater and the consensus in the group (Landis and Koch, 1977a; Soeken and Prescott, 1986; Salerno et al., 2003). Thus, the so-called "consensus

method" reduces the problem to computing the classical Cohen's kappa coefficient. The consensus may be defined as the category chosen by a given proportion of raters in the group (for example, Ruperto et al. (2006) defined the consensus as the category chosen by at least 80% of the raters in the group) or the category the most frequently chosen by the raters in the group (Kalant et al., 2000; Smith et al., 2003). In both cases, however, the problem of handling items without consensus in the group arises. Ruperto et al. (2006) discarded all items without consensus from the analysis, while Kalant et al. (2000) and Smith et al. (2003) did not encounter the problem. The method consisting in reducing the judgements made by a group of raters into a consensus decision was criticized by Eckstein et al. (1998), Salerno et al. (2003) and Miller et al. (2004). Eckstein et al. (1998) studied the bias that may result from removing items without consensus, while Salerno et al. (2003) argued that the dispersion likely to occur in the classifications made by the raters in the group may not be reflected in the consensus. Finally, Miller et al. (2004) showed that different conclusions may be obtained by using different rules of consensus.

Williams (1976) developed a measure for comparing the joint agreement of several raters with another rater without determining a consensus in the group of raters. Specifically, he compared the mean proportion of concordant items between the isolated rater and each rater in the group to the mean proportion of concordant items between all possible pairs of raters among the group. The ratio derived, known as Williams' index, is compared to the value of 1. Unfortunately, Williams' index does neither account for agreement due to chance nor measure the agreement between the isolated rater and the group of raters. In a different context, Schouten (1982) described a hierarchical clustering method based on pairwise weighted agreement measures (referred hereafter as Schouten's agreement index) to identify homogeneous subgroups among a group of raters classifying items on a nominal or ordinal scale. Lastly, Light (1971) investigated the reverse problem of comparing the joint agreement of several raters with a gold standard. He derived a statistic based on the proportion of concordant pairs obtained between each rater in the group and the gold standard (the isolated rater). As for Williams' index, Light's method does not actually quantify the agreement between the gold standard and the group of raters.

Vanbelle and Albert (2009a) proposed a novel coefficient for quantifying the agreement between an isolated rater and a group of raters, considered as a well-defined entity with its own heterogeneity. This coefficient overcomes the problems of consensus by capturing the variability within the group of raters. It generalizes the approach of Schouten (1982) and possesses the same properties as Cohen's kappa coefficient.

4.2 A novel agreement index

4.2.1 Binary scale

Consider a population $\mathcal{I}$ of items and a population $\mathcal{R}$ of raters. Suppose that the items have to be classified in two categories ($K = 2$) by the raters of the population and by an isolated rater, not belonging to $\mathcal{R}$. Consider a randomly selected rater r from population $\mathcal{R}$ and a randomly selected item i from population $\mathcal{I}$. Let $X_{i,r}$ be the random variable such that $X_{i,r} = 1$ if rater r classifies item i in category 1 and $X_{i,r} = 0$ otherwise. For each item i, $E(X_{i,r}|i) = P(X_{i,r} = 1) = P_i$ over the population of raters and $var(X_{i,r}|i) = P_i(1 - P_i)$. Then, over the population of items, $E(P_i) = E[E(X_{i,r}|i)] = \pi$ and $var(P_i) = \sigma^2$. Suppose that the agreement in the population of raters is quantified by the intraclass correlation coefficient (see Chapter 2, Section 2.4.1), labeled ICC in this chapter for convenience reasons,

$$ICC = \frac{\sigma^2}{\pi(1-\pi)}.$$

In the same way, let Y_i denote the random variable equal to 1 if the isolated rater classifies item i in category 1 and $Y_i = 0$ otherwise. Over the population of items, $E(Y_i) = \pi^*$ and $var(Y_i) = \sigma^{*2} = \pi^*(1 - \pi^*)$. The correlation between P_i and Y_i over $\mathcal{I}$ writes

$$\rho = \frac{E(P_iY_i) - \pi\pi^*}{\sigma\sigma^*}.$$

Now, consider the joint probability distribution of the classification of item i made by the population of raters and the isolated rater. On a binary scale, this consists of 4 probabilities $(1 - P_i)(1 - Y_i), (1 - P_i)Y_i, P_i(1 - Y_i)$ and P_iY_i, respectively. For example, P_iY_i denotes the probability that the population of raters and the isolated rater both classify item i in category 1. The expectations, over the population of items, of these joint probabilities can be represented in a 2×2 classification table, as displayed in Table 4.1.

The probability that the population of raters and the isolated rater agree on item i is given by

$$\Pi_i = P_iY_i + (1 - P_i)(1 - Y_i) \tag{4.1}$$

so that, over the population of items $\mathcal{I}$, the mean probability of agreement is given by the expression

$$\Pi_T = E(\Pi_i) = \pi\pi^* + (1 - \pi)(1 - \pi^*) + 2\rho\sigma\sigma^* \tag{4.2}$$

which corresponds to the sum of the diagonal elements in Table 4.1. Surprisingly, for a given level of agreement (ICC) within the population of raters, the maximum attainable value Π_T is not necessarily equal to 1 as shown below.

Table 4.1. Expected joint and marginal probability distributions resulting from the binary classification of a randomly selected item i from the population $\mathcal{I}$ by the population of raters $\mathcal{R}$ and the isolated rater

	Isolated rater		
$\mathcal{R}$	0	1	
0	$E[(1-P_i)(1-Y_i)]$ $(1-\pi)(1-\pi^*)+\rho\sigma\sigma^*$	$E[(1-P_i)Y_i]$ $(1-\pi)\pi^*-\rho\sigma\sigma^*$	$1-\pi$
1	$E[P_i(1-Y_i)]$ $\pi(1-\pi^*)-\rho\sigma\sigma^*$	$E[P_iY_i]$ $\pi\pi^*+\rho\sigma\sigma^*$	π
	$1-\pi^*$	π^*	1

By definition, the population of raters and the isolated rater "perfectly agree" when $\pi = \pi^*$ and $\rho = 1$ (Vanbelle and Albert, 2009a). In terms of the random variables P_i and Y_i, this is equivalent to writing (see proof in Section 4.10.1)

$$P_i = \pi^{**}(1-\sqrt{ICC}) + \sqrt{ICC}\ Y_i.$$

where, for convenience, π^{**} denotes the common value of $\pi = \pi^*$.

Replacing P_i in Equation 4.1 and taking the expectation over population $\mathcal{I}$, the maximum attainable value of Π_T is found to be

$$\Pi_M = 1 - 2\pi^{**}(1-\pi^{**})(1-\sqrt{ICC}). \tag{4.3}$$

This quantity turns out to be equal to 1 if and only if $ICC = 1$, i.e., there is perfect agreement in the population of raters $\mathcal{R}$, or trivially, if $\pi^{**} = 0$ or 1. It should be remarked at this stage that Schouten (1982), in his paper, implicitly assumed $\Pi_M = 1$.

Following the results above, the coefficient of agreement between the population of raters and the isolated rater can be advantageously defined in a kappa-like manner, namely,

$$\kappa = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E} \tag{4.4}$$

with Π_T the theoretical agreement, Π_M the maximum attainable agreement and Π_E the agreement expected by chance. Π_E is the probability that the population of raters and the isolated rater agree under the independence assumption, $E(P_iY_i) = E(P_i)E(Y_i)$. Π_E is defined by

$$\Pi_E = \pi\pi^* + (1-\pi)(1-\pi^*). \tag{4.5}$$

Note that $\Pi_T = \Pi_E$ (see Equations 4.2 and 4.5) in the absence of correlation between the ratings of the population of raters and of the isolated rater ($\rho = 0$) or when there is no variability in the classifications made by the population of raters ($\sigma^2 = 0$) or by the isolated rater ($\sigma^{*2} = 0$). The agreement coefficient (Equation 4.4) has been standardized in such a way that $\kappa = 1$ if the agreement between the isolated rater and the group of raters reaches the maximum attainable value Π_M (perfect agreement) and $\kappa = 0$ when agreement can only be explained by pure chance. Lastly, observe that Equation 4.3 reduces to Schouten's index when $\Pi_M = 1$.

An intraclass version of κ can be derived using the additional assumption $\pi = \pi^* = \pi^{**}$ (equality of marginal probabilities). In that case, we have

$$\kappa_I = \frac{E(P_i Y_i) - \pi^{**2}}{\sigma^{**2}} \tag{4.6}$$

which is equivalent to the correlation coefficient between P_i and Y_i under the assumption of equal marginal probabilities.

4.2.2 Nominal scale

When $K > 2$, the coefficient of agreement between the population of raters and the isolated rater is defined by

$$\kappa = \frac{\sum_{j=1}^{K}(\Pi_{[j]T} - \Pi_{[j],E})}{\sum_{j=1}^{K}(\Pi_{[j]M} - \Pi_{[j]E})} = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E}$$

where $\Pi_{[j]T}$, $\Pi_{[j]E}$ and $\Pi_{[j]M}$ correspond to the quantities described in the binary case ($K = 2$) when the nominal scale is dichotomized by grouping all categories other than category j together. Π_T, Π_E and Π_M are defined respectively by

$$\Pi_T = \sum_{j=1}^{K} E[P_{ij} Y_{ij}];$$

$$\Pi_E = \sum_{j=1}^{K} \pi_j \pi_j^*;$$

$$\Pi_M = \sum_{j=1}^{K} E[(\pi_j^{**} + (1 - \pi_j^{**})\sqrt{ICC_j})Y_{ij}] = \sum_{j=1}^{K}(\pi_j^{**} + \pi_j^{**}(1 - \pi_j^{**})\sqrt{ICC_j})$$

where P_{ij} denotes the probability for a randomly selected item i to be classified in category j ($j = 1, \cdots, K$) by the population of raters, with $E(P_{ij}) = \pi_j$. Y_{ij} denotes the random variable equal to 1 if the isolated rater classifies item i in category j ($Y_{ij} = 0$ otherwise). Finally, ICC_j denotes the intraclass kappa coefficient relative to category j ($j = 1, \cdots, K$) in the population of raters (see proof

in Section 4.10.2).

The coefficient κ possesses the same properties as Cohen's kappa coefficient, $\kappa = 1$ when agreement is perfect ($\Pi_T = \Pi_M$), $\kappa = 0$ if observed agreement is equal to agreement expected by chance ($\Pi_T = \Pi_E$) and $\kappa < 0$ if observed agreement is lower than expected by chance ($\Pi_T < \Pi_E$).

4.2.3 Ordinal scale

A weighted version of the agreement index can be defined in a way similar to the weighted kappa coefficient (see Chapter 2, Section 2.5),

$$\kappa_W = \frac{\Pi_{T,W} - \Pi_{E,W}}{\Pi_{M,W} - \Pi_{E,W}}$$

with

$$\Pi_{T,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E(P_{ij} Y_{ik});$$

$$\Pi_{E,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} \pi_j \pi_k^*;$$

$$\Pi_{M,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E[(\pi_j^{**} + (1-\pi_j^{**})\sqrt{ICC_j} Y_{ij}) Y_{ik}].$$

4.3 Estimation of the parameters

Consider a random sample of N items drawn from population $\mathcal{I}$. Let each item be classified independently on a K-categorical scale by a random sample (group) of R raters from population $\mathcal{R}$ and by the isolated rater.

4.3.1 Binary scale

Let $x_{i,r}$ designate the observed value of the random variable $X_{i,r}$, denoting the category assignment made for item i by rater r from population $\mathcal{R}$ ($i = 1, \cdots, N; r = 1, \cdots, R$). Then, let $n_i = \sum_{r=1}^{R} x_{i,r}$ denote the number of times that item i is classified in category 1 by the group of raters and $p_i = n_i/R$ the corresponding proportion ($i = 1, \cdots, N$). If y_i denotes the observed value of the random variable Y_i, representing the category assignment of item i by the isolated rater, the probability that the population of raters and the isolated rater agree is estimated by the *observed proportion of agreement*,

$$p_o = \widehat{\Pi}_T = \frac{1}{N}\sum_{i=1}^{N} \left[p_i y_i + (1-p_i)(1-y_i)\right]. \tag{4.7}$$

The probability of agreement expected by chance is estimated by the *proportion of agreement expected by chance*,

$$p_e = \widehat{\Pi}_E = py + (1-p)(1-y)$$

where y is the proportion of items classified in category 1 by the isolated rater,

$$y = \frac{1}{N}\sum_{i=1}^{N} y_i$$

and p is the overall proportion of items classified in category 1 by the group of raters,

$$p = \frac{1}{N}\sum_{i=1}^{N} p_i.$$

The degree of agreement κ between the group of raters and the isolated rater is then estimated by

$$\widehat{\kappa} = \frac{p_o - p_e}{p_m - p_e}$$

where p_m corresponds to the maximum possible proportion of agreement derived from the sample. Since each response y_i given by the isolated rater can only be 0 or 1, it is easily seen that for each item i, $p_i y_i + (1-p_i)(1-y_i) \leq max(p_i, 1-p_i)$ $(i = 1, \cdots, N)$. It follows from Equation 4.7 that the maximum attainable proportion of agreement is given by the expression

$$p_m = \widehat{\Pi}_M = \frac{1}{N}\sum_{i=1}^{N} max(p_i, 1-p_i).$$

This quantity can only be equal to 1 if $p_i = 0$ or 1 for all items $(i = 1, \cdots, N)$ as assumed by Schouten.

4.3.2 Nominal scale

The estimation of the parameters easily extends to the case $K > 2$. Let $x_{ij,r}$ denote the observed value of the random variable $X_{ij,r}$ equal to 1 if rater r $(r = 1, \cdots, R)$ of the group classifies item i $(i = 1, \cdots, N)$ in category j $(j = 1, \cdots, K)$ and equal to 0 otherwise. Then, let $n_{ij} = \sum_{r=1}^{R} x_{ij,r}$ denote the number of times the item i is classified in category j by the raters of the group and p_{ij} the corresponding proportion. We have $\sum_{j=1}^{K} p_{ij} = 1$, $(i = 1, \cdots, N)$. Finally, let y_{ij} denote the observed value of the random variable Y_{ij} corresponding to the category assignment of item i made by the isolated rater. Then, the data can be conveniently summarized in a two-way classification table (see Table 4.2) by defining the quantities

$$c_{jk} = \frac{1}{N}\sum_{i=1}^{N} p_{ij} y_{ik}, \;\; j, k = 1, \cdots, K.$$

Table 4.2. Two-way classification table of the N items by the group of raters and the isolated rater

	Isolated rater					
Group of raters	1	...	j	...	K	Total
1	c_{11}	...	c_{1j}	...	c_{1K}	$c_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	c_{j1}	...	c_{jj}	...	c_{jK}	$c_{j.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	c_{K1}	...	c_{Kj}	...	c_{KK}	$c_{K.}$
Total	$c_{.1}$	...	$c_{.j}$	...	$c_{.K}$	1

The *observed proportion of agreement* between the group of raters and the isolated rater is defined by

$$p_o = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}y_{ij} = \sum_{j=1}^{K} c_{jj}.$$

The marginal classification distribution of the isolated rater, namely,

$$y_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij}, \quad j = 1, \cdots, K \tag{4.8}$$

with $\sum_{j=1}^{K} y_j = 1$ and the marginal classification distribution of the group of raters,

$$p_j = \frac{1}{N}\sum_{i=1}^{N} p_{ij}, \ j = 1 \cdots, K \tag{4.9}$$

with $\sum_{j=1}^{K} p_j = 1$ are needed to estimate the agreement expected by chance. The *proportion of agreement expected by chance* is given by

$$p_e = \sum_{j=1}^{K} p_j y_j = \sum_{j=1}^{K} c_{j.}c_{.j}.$$

The degree of agreement κ between the population of raters and the isolated rater is then estimated by

$$\widehat{\kappa} = \frac{p_o - p_e}{p_m - p_e}$$

where p_m corresponds to the maximum possible proportion of agreement derived from the data set. By extending the argument used for the binary case, it is easily seen that

$$p_m = \frac{1}{N}\sum_{i=1}^{N} max_j p_{ij}. \tag{4.10}$$

Observe that in the calculation of p_m, no explicit use is made of category j in which the maximum occurs. Thus, in case where the maximum is not unique, only the value of the maximum is actually important.

4.3.3 Ordinal scale

The estimation of the weighted agreement index is simply done by introducing weights in the estimations previously defined. Hence,

$$\widehat{\kappa}_W = \frac{p_{o,w} - p_{e,w}}{p_{m,w} - p_{e,w}}$$

with

$$\begin{aligned} p_{o,w} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}p_{ij}y_{ik} \\ p_{e,w} &= \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}p_j y_k \\ p_{m,w} &= \frac{1}{N}\sum_{i=1}^{N} max_j(\sum_{k=1}^{K} w_{jk}p_{ik}). \end{aligned}$$

The unweighted agreement index $\widehat{\kappa}$ can be obtained using the weights $w_{jj} = 1$ and $w_{jk} = 0,\ j \neq k$.

4.3.4 Sampling variability

The Jackknife method (Efron and Tibshirani, 1993) can be used to determine the sampling variance of the agreement index. Suppose that the agreement between the isolated rater and the population of raters was estimated on a random sample of N items. Let $\widehat{\kappa}_N$ denote that agreement index and $\widehat{\kappa}_{N-1}^{(i)}$ the estimated agreement index when observation i is deleted. These quantities are used to determine the pseudo-values

$$\widehat{\kappa}_{N,i} = N\widehat{\kappa}_N - (N-1)\widehat{\kappa}_{N-1}^{(i)}.$$

The Jackknife estimator of the agreement index is then defined by

$$\widetilde{\kappa}_N = \frac{1}{N}\sum_{i=1}^{N}\widehat{\kappa}_{N,i}$$

with variance

$$var(\widetilde{\kappa}_N) = \frac{1}{N}\left\{\frac{1}{N-1}\sum_{i=1}^{N}(\widehat{\kappa}_{N,i} - \widehat{\kappa}_N)^2\right\}.$$

The bias of the Jackknife estimator is estimated by

$$Bias(\widetilde{\kappa}_N) = (N-1)\left\{\widetilde{\kappa}_N - \widehat{\kappa}_N\right\}.$$

4.3.5 Example

Consider the following hypothetic example (Vanbelle et al., 2007) to illustrate how to calculate the proposed agreement index. Suppose that an isolated rater and a group of 12 raters have to classify 3 items on a 5-point Likert scale with values -2, -1, 0, 1 and 2. The data are given in Table 4.3.

Table 4.3. Classification of 3 items on a 5-point Likert scale given by a group of 12 raters and an isolated rater (hypothetic example)

	Raters in the group												Isolated rater
Item	1	2	3	4	5	6	7	8	9	10	11	12	1
1	0	1	2	2	2	1	2	1	1	1	1	1	1
2	0	-1	1	0	0	-1	-1	0	0	-1	-1	-1	0
3	1	1	-2	-1	-1	1	-2	-2	-1	-1	1	1	-2

The responses given by the group of raters can then be summarized (see Table 4.4). For example, 7 raters of the group have classified item 1 in category (1).

Table 4.4. Distribution of the responses given by the group of 12 raters and the isolated rater (hypothetic example)

	Group of raters					Isolated rater				
	Category					Category				
Item	(-2)	(-1)	(0)	(1)	(2)	(-2)	(-1)	(0)	(1)	(2)
1	0	0	1	7	4	0	0	0	1	0
2	0	6	5	1	0	0	0	1	0	0
3	3	4	0	5	0	0	1	0	0	0

The responses of the group of raters and the isolated rater are then expressed in terms of proportions ($p_{ij} = n_{ij}/12$) and the marginal classification distribution of the group of raters (p_j) determined using Equation 4.9. In the same way, the marginal distribution of the isolated rater (y_j) can be determined by Equation 4.8. The values of these parameters are given in Table 4.5.

The observed proportion of agreement is equal to

$$p_o = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}y_{ij} = \frac{0.58 + 0.42 + 0.25}{3} = 0.42,$$

Table 4.5. Distribution of the responses given by the group of 12 raters and the isolated rater expressed in terms of proportion (hypothetic example)

	Group of raters					Isolated rater				
	Category					Category				
Item	(-2)	(-1)	(0)	(1)	(2)	(-2)	(-1)	(0)	(1)	(2)
1	0	0	0.08	0.58	0.33	0	0	0	1	0
2	0	0.50	0.42	0.08	0	0	0	1	0	0
3	0.25	0.33	0	0.42	0	1	0	0	0	0
p_j	0.08	0.28	0.17	0.36	0.11	0.33	0	0.33	0.33	0

the proportion of agreement expected by chance to

$$p_e = \sum_{j=1}^{K} p_j y_j = 0.08 \times 0.33 + 0.28 \times 0 + 0.17 \times 0.33 + 0.36 \times 0.33 + 0.11 \times 0 = 0.20$$

and the maximum possible proportion of agreement to

$$p_m = \frac{1}{N} \sum_{i=1}^{N} max_j p_{ij} = \frac{0.58 + 0.50 + 0.42}{3} = 0.50.$$

This leads to an agreement index of

$$\widehat{\kappa} = \frac{p_o - p_e}{p_m - p_e} = \frac{0.42 - 0.20}{0.50 - 0.20} = 0.73.$$

4.4 The consensus approach

4.4.1 Binary scale

Consider a population $\mathcal{I}$ of items and a population $\mathcal{R}$ of raters. Suppose that the items have to be classified in two categories ($K = 2$) by the raters of the population and by an isolated rater, not belonging to $\mathcal{R}$. As already mentioned, the consensus approach consists in summarizing the responses given by the raters of the group in a unique quantity for each item. Very often the consensus category is taken as the modal category (majority rule) or the category chosen by a prespecified proportion of raters (e.g., $\geq 80\%$). Evidently, a consensus may not always be defined. For example, on a nominal scale, one could have two modal categories or no category chosen by the prespecified proportion of raters. Therefore, suppose that on the N items drawn from population $\mathcal{I}$, a consensus can only be defined on $N_C \leq N$ items. Let $\mathcal{I}_C$ denote the sub-population of items on which a consensus is always possible and Z_i be a random variable equal to 1 if category 1 corresponds to the consensus category given by the population $\mathcal{R}$ of raters for item i and equal to 0

otherwise. Then, over $\mathcal{I}_C$, $E(Z_i) = \phi$ and $var(Z_i) = \sigma_C^2 = \phi(1-\phi)$. In the same way, let Y_i denote the random variable equal to 1 if the isolated rater classifies item i in category 1 and $Y_i = 0$ otherwise. Over the population of items, $E(Y_i) = \pi^*$ and $var(Y_i) = \sigma^{*2} = \pi^*(1-\pi^*)$. If ρ' denotes the correlation coefficient between Y_i and Z_i, we have the following representation of the cross-classifications of the items by the isolated rater and the population of raters (Table 4.6).

Table 4.6. Expected probabilities of the classification of the isolated rater and the population of raters over the sub-population $\mathcal{I}_C$ of items where a consensus is possible

		Isolated rater		
		0	1	
$\mathcal{R}$	0	$E((1-Z_i)(1-Y_i))$ $(1-\phi)(1-\pi^*) + \rho'\sigma_C\sigma^*$	$E((1-Z_i)Y_i)$ $(1-\phi)\pi^* - \rho'\sigma_C\sigma^*$	$1-\phi$
	1	$E(Z_i(1-Y_i))$ $\phi(1-\pi^*) - \rho'\sigma_C\sigma^*$	$E(Z_iY_i)$ $\phi\pi^* + \rho'\sigma_C\sigma^*$	ϕ
		$1-\pi^*$	π^*	1

The agreement between the consensus in the population of raters and the isolated rater is defined by

$$\Pi_{iC} = Z_iY_i + (1-Z_i)(1-Y_i). \tag{4.11}$$

Thus,

$$\Pi_{TC} = E(\Pi_{iC}) = \phi\pi^* + (1-\phi)(1-\pi^*) + 2\rho'\sigma_C\sigma^*. \tag{4.12}$$

The agreement expected by chance is defined by

$$\Pi_{EC} = \phi\pi^* + (1-\phi)(1-\pi^*) \tag{4.13}$$

and perfect agreement is achieved when $Z_i = Y_i$, for all items in $\mathcal{I}_C$, leading to

$$\Pi_{MC} = E(\Pi_i) = 1. \tag{4.14}$$

Therefore, the agreement coefficient between the population of raters and the isolated rater is defined by

$$\kappa_C = \frac{\Pi_{TC} - \Pi_{EC}}{1 - \Pi_{EC}} \tag{4.15}$$

and corresponds to Equation 2.18 derived in Chapter 2, Section 2.3.3.

4.4.2 Nominal scale

Equation 4.15 can be extended to the case of a scale with $K > 2$ categories in the following way:

$$\kappa_C = \frac{\sum_{j=1}^{K}(\Pi_{[j]TC} - \Pi_{[j]EC})}{\sum_{j=1}^{K}(1 - \Pi_{[j]EC})} = \frac{\Pi_{TC} - \Pi_{EC}}{1 - \Pi_{EC}} \tag{4.16}$$

where $\Pi_{[j]TC}$ and $\Pi_{[j]EC}$ correspond to the quantities described in the binary case when the nominal scale is dichotomized by grouping all categories other than category j together and Π_T, Π_E and Π_M are defined respectively by

$$\Pi_{TC} = \sum_{j=1}^{K} E(Z_{ij}Y_{ij}) \text{ and } \Pi_{EC} = \sum_{j=1}^{K} \phi_j \pi_j^* \tag{4.17}$$

where $Z_{ij} = 1$ if category j corresponds to the consensus in the population of raters for item i ($Z_{ij} = 0$ otherwise) with $E(Z_{ij}) = \phi_j$ and $Y_{ij} = 1$ if item i is classified by the isolated rater in category j ($Y_{ij} = 0$ otherwise) with $E(Y_{ij}) = \pi_j^*$.

4.4.3 Ordinal scale

The weighted version of the agreement index (κ_{WC}) can the be derived by introducing weights in the expression of Π_{TC} and Π_{EC} in the following way,

$$\Pi_{T,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E(Z_{ij}Y_{ik}) \text{ and } \Pi_{E,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}\phi_j \pi_k^* \tag{4.18}$$

leading to the agreement index

$$\kappa_{WC} = \frac{\Pi_{T,WC} - \Pi_{E,WC}}{1 - \Pi_{E,WC}}. \tag{4.19}$$

4.4.4 Estimation of the parameters

Suppose that z_{ij} (resp. y_{ij}) denote the observed values of the random variables Z_{ij} (resp. Y_{ij}) ($i = 1, \cdots, N_C$) defined in Section 4.4.2. The assessment of the N_C items on which it is possible to determine a consensus by the two groups of raters can be conveniently summarized by the quantities

$$d_{jk} = \frac{1}{N_C}\sum_{i=1}^{N_C} z_{ij}y_{ik} \ (j, k = 1, \cdots, K). \tag{4.20}$$

Similarly to what was done in Section 4.3, the *observed weighted agreement* between the two groups of raters is obtained by

$$\widehat{\Pi}_{T,WC} = p_{o,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} d_{jk} = \frac{1}{N_C}\sum_{i=1}^{N_C}\sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} z_{ij} y_{ik} \tag{4.21}$$

and the *weighted agreement expected by chance* by the expression

$$\widehat{\Pi}_{E,WC} = p_{e,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} d_{j.} d_{.k} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} z_j y_k \tag{4.22}$$

where $z_j = \frac{1}{N_C}\sum_{i=1}^{N_C} z_{ij}$. This leads to the agreement index

$$\hat{\kappa}_{WC} = \frac{p_{o,WC} - p_{e,WC}}{1 - p_{e,WC}}. \tag{4.23}$$

4.4.5 Example

Let illustrate the consensus method on the example developed in Section 4.3.5. Using the majority rule (consensus category = category chosen by the majority of the raters) to determine a consensus in the group of raters, the consensus category corresponds to category (1), (-1) and (1) for items 1, 2 and 3, respectively. This results were cross-classified with the responses given by the isolated rater to be summarized in a 5×5 contingency Table (see Table 4.7).

Table 4.7. Cross-classification of the responses given by the group of raters (consensus) and the isolated rater on 3 items in terms of proportion

		Isolated rater					
	Category	(-2)	(-1)	(0)	(1)	(2)	Total
Group of raters	(-2)	0	0	0	0	0	0
	(-1)	0	0	0.33	0	0	0.33
	(0)	0	0	0	0	0	0
	(1)	0	0	0	0.33	0.33	0.66
	(2)	0	0	0	0	0	0
	Total	0	0	0.33	0.33	0.33	1

The observed proportion of agreement is equal to

$$p_{oC} = \sum_{j=1}^{K} d_{jj} = 0.33,$$

and the proportion of agreement expected by chance to

$$p_{eC} = \sum_{j=1}^{K} d_{j\cdot} d_{\cdot j} = 0 \times 0 + 0 \times 0.33 + 0.33 \times 0 + 0.33 \times 0.66 + 0.33 \times 0 = 0.22.$$

Cohen's kappa coefficient is then equal to

$$\widehat{\kappa}_C = \frac{p_{oC} - p_{eC}}{1 - p_{eC}} = \frac{0.33 - 0.22}{1 - 0.22} = 0.14.$$

4.5 Schouten's agreement index

Schouten (1982) derived an index to select one or more homogeneous subgroups of raters when each item of a sample of items is classified on a K-category scale by each of a fixed group of $R+1$ raters. In this perspective, Schouten (1982) introduced weighted agreement indexes to measure the degree of agreement between two particular raters, between a particular rater and the other raters of the group and within subgroups of raters.

4.5.1 Definition

Suppose that N items were classified in K categories by a group of $R+1$ raters. Let $p_{r,s}(j,k)$ denote the proportion of items assigned in category j by rater r and category k by rater s ($j,k \in \{1,\cdots,K\}$;$r,s \in \{1,\cdots,R+1\}$). The proportion

$$p_r(j,k) = \frac{1}{R}\sum_{s\neq r} p_{r,s}(j,k) \tag{4.24}$$

was introduced by Schouten (1982) to estimate the probability that a randomly selected item is assigned to category j by rater r and to category k by an randomly taken rater from the remaining R raters. Then, Schouten (1982) defined the proportion

$$p(j,k) = \frac{1}{R(R+1)}\sum_{r=1}^{R+1}\sum_{s\neq r} p_{r,s}(j,k) \tag{4.25}$$

to estimate the probability for a randomly selected item to be assigned to category j and k by two raters randomly taken for the population of raters.

Finally, Schouten (1982) denoted by

$$p_r(j) = \sum_{k=1}^{K} p_{r,s}(j,k) \tag{4.26}$$

the proportion of items assigned to category j by rater r. The proportion

$$q_{r,s}(j,k) = p_r(j)p_s(k) \tag{4.27}$$

then estimated to probability that for a randomly selected item to be assigned in category j by rater r and category k by rater s if the two assignments were independently distributed. Then,

$$q_r(j,k) = \frac{1}{R}\sum_{s\neq r} q_{r,s}(j,k) \tag{4.28}$$

estimates the probability that a randomly selected item is assigned to category j by the rater r and to category k by another rater taken randomly from the remaining R raters. Finally,

$$q(j,k) = \frac{1}{R(R+1)} \sum_{r=1}^{R+1} \sum_{s \neq r} q_{r,s}(j,k) \tag{4.29}$$

is an estimate of the probability that a randomly selected item is assigned to category j by the first and to category k by the second of 2 raters who are taken at random and without replacement from the whole group of raters.

For two raters r and s, Schouten (1982) defined the *observed weighted agreement* by

$$o_{r,s}(w) = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{r,s}(j,k) \tag{4.30}$$

and the weighted agreement expected by chance by

$$e_{r,s}(w) = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} q_{r,s}(j,k) \tag{4.31}$$

leading to a weighted kappa coefficient between raters r and s of

$$\hat{\kappa}_{r,s}(w) = \frac{o_{r,s}(w) - e_{r,s}(w)}{1 - e_{r,s}(w)}. \tag{4.32}$$

This corresponds to the usual definition of the weighted kappa coefficient (Cohen, 1968) between two single raters (see Chapter 2, Section 2.5).

The measure of agreement between rater r and the other R raters of the group was defined by Schouten (1982) to be

$$\hat{\kappa}_r(w) = \frac{o_r(w) - e_r(w)}{1 - e_r(w)} \tag{4.33}$$

where

$$o_r(w) = \frac{1}{R} \sum_{s \neq r} o_{r,s}(w) \tag{4.34}$$

and

$$e_r(w) = \frac{1}{R} \sum_{s \neq r} e_{r,s}(w). \tag{4.35}$$

Finally, Schouten (1982) defined the weighted kappa coefficient $\kappa(w)$ as a group measure of agreement among the $R+1$ raters:

$$\hat{\kappa}(w) = \frac{o(w) - e(w)}{1 - e(w)} \tag{4.36}$$

where

$$o(w) = \frac{1}{R(R+1)} \sum_{r=1}^{R+1} \sum_{s \neq r} o_{r,s}(w) \tag{4.37}$$

and

$$e(w) = \frac{1}{R(R+1)} \sum_{r=1}^{R+1} \sum_{s \neq r} e_{r,s}(w). \tag{4.38}$$

Using the agreement weights $w_{jj} = 1$ and $w_{jk} = 0$ $(j, k = 1, \cdots, K)$, the weighted kappa coefficient is equivalent to the pairwise agreement index derived by Davies and Fleiss (1982) (see Chapter 3, Section 3.3.2).

4.5.2 Example

Consider the data in Table 4.3. For simplicity, consider the weights $w_{jj} = 1$ and $w_{jk} = 0$ $(j, k = 1, \cdots, K)$. The observed proportion of agreement and the proportion of agreement expected by chance between each rater of the group and the isolated rater are given in Table 4.8. This leads to a Schouten's agreement index of

$$\hat{\kappa}_r(w) = \frac{o_r(w) - e_r(w)}{1 - e_{r(w)}} = \frac{0.42 - 0.20}{1 - 0.20} = 0.27.$$

Table 4.8. Proportion of observed agreement $o_{r,s}(w)$, of agreement expected by chance $e_{r,s}(w)$ and Cohen's kappa coefficient $\hat{\kappa}_r(w)$ between the isolated rater and each rater of the group

Rater	1	2	3	4	5	6	7	8	9	10	11	12
$o_{r,s}(w)$	0.33	0.33	0.33	0.33	0.33	0.33	0.33	1	0.67	0.33	0.33	0.33
$e_{r,s}(w)$	0.33	0.22	0.22	0.11	0.11	0.22	0.11	0.33	0.22	0.11	0.22	0.22
$\hat{\kappa}_r(w)$	0	0.14	0.14	0.25	0.25	0.14	0.25	1	0.57	0.25	0.14	0.14

The mean of the kappa coefficients is equal to 0.273, while Schouten's index amounts 0.267. Remark that these values are close but not equal.

4.6 William's agreement index

4.6.1 Definition

The idea of Williams (1976) was to derive an agreement index giving an answer to the following question: given a group of raters (namely, raters $1, \cdots, R$) and one other rater (rater $R+1$), does the isolated rater agree with the group of raters as

often as a member of that group agrees with another member in the group? Using the notation introduced in Section 4.5, William's agreement index is

$$\hat{I}_R = \frac{o_{R+1}(w)}{o(w)} \tag{4.39}$$

with

$$o(w) = \frac{1}{R(R-1)} \sum_{r=1}^{R} \sum_{s \neq r} o_{r,s}(w). \tag{4.40}$$

Then, Williams (1976) used Normal approximation to test if the ratio $\hat{I}_R$ is different from the value 1, in which case the rate of agreement obtained between the isolated rater and the group of raters is different from the rate of agreement in the group of raters.

4.6.2 Example

Consider again the data in Table 4.3 with weights $w_{jj} = 1$ and $w_{jk} = 0$ $(j, k = 1, \cdots, K)$. The observed proportion of agreement between the isolated rater and the raters in the group is calculated in the same way as for Schouten's agreement index and is equal to $o_r(w) = 0.42$. Since the observed agreement in the group of raters is equal to $o(w) = 0.36$, William's agreement index is equal to $\hat{I}_R = 0.42/0.36 = 1.17$.

4.7 Comparison of the agreement indexes

4.7.1 Comparison with the consensus method

There are two major differences between the consensus method and the agreement index proposed by Vanbelle and Albert (2009a). Firstly, a consensus method can not always be defined while the new agreement index can always be determined. For example, using the majority rule, there is no consensus in the group of raters if the distribution of the responses are uniformly distributed. Secondly, the strength of the consensus is not taken into account by the random variable Z_{ij} while the proposed agreement does, being based on the probability distribution of the responses in the group of raters. For example on a binary scale, using the majority rule, we will have $Z_{ij} = 1$ if $P_{ij} = 0.6$ but also if $P_{ij} = 0.9$.

It can easily be shown that the new methodology defined by Vanbelle and Albert (2009a) and the consensus approach are equivalent only in two particular cases, firstly when there is only one rater in the group of raters ($R = 1$) and secondly when $\mathcal{I}_C = \mathcal{I}$ and there is perfect agreement in the population of raters ($ICC = 1$).

4.7.2 Comparison with Schouten's index

We can easily show that $p_{o,w}$ and $o_{R+1}(w)$ are equivalent. Indeed,

$$\begin{aligned} p_{o,w} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}p_{ij}y_{ik} \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}\frac{1}{R}\sum_{r=1}^{R} x_{ij,r}y_{ik} \\ &= \frac{1}{R}\sum_{r=1}^{R}\sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}\frac{1}{N}\sum_{i=1}^{N} x_{ij,r}y_{ik} \\ &= \frac{1}{R}\sum_{r=1}^{R}\sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}p_{r,R+1}(j,k) \\ &= \frac{1}{R}\sum_{r=1}^{R} o_{r,R+1}(j,k) \\ &= o_{R+1}(j,k). \end{aligned} \tag{4.41}$$

In the same way, $p_{e,w}$ and $e_{R+1}(w)$ are equivalent. The difference between the agreement index of Schouten (1982) and the agreement index proposed by Vanbelle and Albert (2009a) lies in the definition of perfect agreement. The definition taken by Schouten is more restrictive, requiring $ICC = 1$ in the population of raters (perfect agreement within the population of raters) to have perfect agreement between the isolated rater and the group of raters.

4.8 Examples

4.8.1 Syphilis serology

In Chapter 3, the syphilis serology example was introduced. 28 syphilis specimens were categorized in 3 categories by 3 reference laboratories and a participant. The agreement between the 3 references laboratories was determined. Discordances occurred between the 3 reference laboratories for seven specimens. Now, let determine the agreement between the participant and the 3 reference laboratories. Data are therefore summarized in a two-way classification table (Table 4.9) as explained in Section 4.3. In this example $R = 3$, $K = 3$ and $N = 28$. Results are summarized in Table 4.10. The standard error was determined with the Jackknife method.

Using the quadratic weighting scheme, the weighted coefficient of agreement $\widehat{\kappa}_W$ ($\pm$ SE) between the participant and the 3 reference laboratories, as defined in Section 4.2, was equal to 0.79 ($\pm$0.06). When applying the consensus approach based on

Table 4.9. Two-way classification table of the 28 syphilis serology specimens as NR (non-reactive), BL (borderline) and RE (reactive) by 3 reference laboratories and participant L

	Participant L			
Reference laboratories	NR	BL	RE	Total
NR	0.143	0.250	0.024	0.417
BL	0	0.036	0.071	0.107
RE	0	0	0.476	0.476
Total	0.143	0.286	0.571	1

Table 4.10. Weighted ($\hat{\kappa}_W$) and unweighted ($\hat{\kappa}$) agreement indexes corresponding the the syphilis serology example

Method	N	$\hat{\kappa}_W \pm SE$	$\hat{\kappa} \pm SE$
Vanbelle and Albert (2009a)	28	0.79±0.06	0.55±0.10
Consensus (majority)	26	0.76±0.06	0.42±0.11
Schouten (1982)	28	0.73±0.07	0.46±0.09

the majority rule, we found a weighted kappa coefficient of 0.76 (±0.06), but two specimens were eliminated because no consensus could be reached between the 3 reference laboratories. The weighted agreement index developed by Schouten (1982) amounted 0.73 (±0.07), while the intraclass kappa coefficient ($I\hat{C}C$) in the reference laboratory group was 0.68 (±0.06). Because of the lack of perfect agreement among the reference laboratories ($I\hat{C}C < 1$), Schouten's agreement index can never be equal to 1 so that perfect agreement can never be attained. According to Equation 4.10, the non-weighted maximum attainable proportion was $p_m = 0.893$, while the corresponding value for the quadratic weighting scheme was $p_{m,w} = 0.973$. To derive the highest possible value of the proposed agreement index, consider the hypothetical laboratory H whose responses are given in Table A.2. For this particular laboratory, since each specimen's result corresponds to the most frequent response given by the reference laboratories, our agreement index yield the perfect value of 1 (±0), while Schouten's index is only equal to 0.94 (±0.025). For the consensus approach, the kappa coefficient derived was also equal to 1, although 2 specimens (16 and 17) have to be excluded. Finally, it should be remarked that if the hypothetical laboratory H had supplied results different from BL for specimens 16 and 17, the non weighted agreement coefficient obtained would still be 1 but the weighted version would yield a value less than 1 because of the weighting scheme ($\hat{\kappa}_W = 0.958$).

4.8.2 Script Concordance Test

The Script Concordance Test (SCT) is used in medicine to evaluate the ability of physicians or medical students (isolated raters) to solve clinical situations not clearly defined (Charlin et al., 2002). The complete test consists of a number of items $(1, \cdots, N)$ to be evaluated on a 5-point Likert scale ($K = 5$). Each item represents a clinical situation likely to be seen in real life practice and a potential assumption is proposed with it. The situation has to be unclear, even for an expert. The task of the student or the physician being evaluated is to consider the effect of additional evidence on the suggested assumption. In this respect, the candidate has to choose between the following proposals: (-2) The assumption is practically eliminated; (-1) The assumption becomes less likely; (0) The information has no effect on the assumption; (+1) The assumption becomes more likely; (+2) The assumption is basically the only possible one. The questionnaire is also given to a panel of experts (raters $1, \cdots, R$). The problem is to evaluate the agreement between each individual medical student and the panel of experts.

Between 2003 and 2005, an SCT was proposed to students training in general practice (Vanbelle et al., 2007). The SCT consisted of 34 items relating possible situations encountered in general practice. There were 39 students passing the test and completing the entire questionnaire. Their responses were confronted to the responses of a panel of 11 experts. The intraclass correlation coefficient in the group of experts was 0.22 ($\pm$0.04). The individual $\widehat{\kappa}_W$ coefficients for the 39 students were computed using the quadratic weighting scheme. Values ranged between 0.37 and 0.84 and the mean agreement index $\pm$ standard deviation (SD) was 0.61 $\pm$ 0.12. Schouten's weighted index scores averaged 0.44 $\pm$ 0.08 (range: 0.26-0.58).

Using the consensus method, where consensus was defined as either the majority of the raters or a proportion of at least 50% of the raters, respectively 2 (6%) and 12 (35%) items had to be omitted from the analysis because no consensus was reached among the experts. The mean weighted kappa values for the 39 students was equal to 0.49 $\pm$ 0.13 (range: 0.19-0.72) with the majority rule and 0.66 $\pm$ 0.14 (range: 0.23-0.82) with the 50% rule. Figure 4.1 displays the individual agreement coefficients relative to each student for the various methods. Marked differences can be seen on the graph depending on the approach used. A ranking of the students was needed for selection purposes. The ranking changed notably for some students according to the agreement index calculated. For example, student No. 39 ranked at the 16th place with the new approach, the 9th place with Schouten index, the 10th place using the majority rule and at 20th place using the 50% rule.

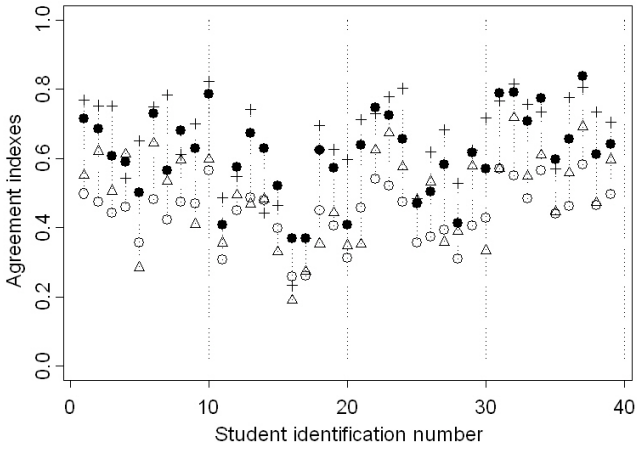

Figure 4.1. Values of $\hat{\kappa}_W$ (•), weighted kappa coefficients using the majority (△) and the 50% (+) rules and weigthed agreement index of Schouten (∘) for the 39 students passing the SCT

4.9 Discussion

Vanbelle and Albert (2009a) developed a method to quantify the agreement between an isolated rater and a group of raters judging items on a categorical scale. The group of raters is seen as a well-defined entity, a reference or gold standard group with its own heterogeneity, whereas the isolated rater comes from a distinct population. Therefore, the marginal classification probabilities of the isolated rater and of the population of raters were basically assumed to be different ($\pi \neq \pi^*$). In the SCT example, it is realistic to admit that each student differs from the group of experts by the knowledge he/she acquired so far in clinical decision-making. Although the group of raters was seen as the "reference" group in the present chapter, the theory is equally applicable to the case where the isolated rater represents the expert, at least as long as a single agreement index is looked for between them. When neither the isolated rater nor the group of raters is considered as the gold standard, an intraclass version of the proposed agreement index can be derived. The latter reduces to the intraclass kappa coefficient (Kraemer, 1979) in case of two isolated raters, by assuming that the isolated rater and the group of raters come from the same population ($\pi = \pi^*$).

The new agreement index was conveniently developed on a population-based model, allowing an easy extension from dichotomous to nominal scales and the use of weighted agreement coefficients. It also leads to a less restrictive definition of perfect agreement. Indeed, the isolated rater and the group of raters were defined to

be in "perfect agreement" when their respective classifications of items were linearly related and equal on average, without perfect agreement among all raters in the group ($ICC < 1$). It was shown that under this assumption and the additional assumption of perfect agreement within the population of raters ($ICC = 1$), the agreement index κ proposed by Vanbelle and Albert (2009a) is algebraically equivalent to the agreement coefficient derived by Schouten (1982). In other terms, the approach of Vanbelle and Albert (2009a) is based on less stringent assumptions than those made by Schouten. This was illustrated on the syphilis example where it was not possible for Schouten's agreement index to achieve the maximum value of 1 to the contrary of the new agreement index. The latter further overcomes the shortcomings of the widely used consensus method, in particular the fact that a decision is not required for items lacking a consensus in the group. It should be remarked, however, that for items lacking consensus among the members of the group, the responses given by the isolated rater can lead to different kappa values depending on the scheme uses (weighted or non weighted) as demonstrated by the hypothetical laboratory in Williams' example. The agreement index proposed by Vanbelle and Albert (2009a) also takes into account the existing variability in the group of raters while the strength of consensus, as already indicated, is completely ignored in the consensus method. Lastly, as illustrated in the SCT example and pointed out by Salerno et al. (2003) and Miller et al. (2004), the results may vary markedly according to the definition of the consensus method used.

The notion of perfect agreement appears to play a major role in the definition of the new agreement coefficient and particularly of its maximum value of 1. Here, the population of raters is seen as a whole, a single entity composed of equally valued members but displaying heterogeneity in their judgments of items. Hence, perfect agreement is defined between the isolated rater and the population of itself, not between the isolated rater and the individual members of the population. As a consequence, agreement may be perfect without forcing all raters, including the isolated one, to classify all items in the same way. The present definition also does not preclude that the agreement between the isolated rater and the population may be better than the agreement between the population and some of its individual members. In other terms, the isolated rater can perform better than some of the experts. This may sound somewhat contradictory in the context of a gold standard. In Schouten's view, an agreement value of 1 can only be achieved when all raters of the population and the isolated rater perfectly and thoroughly agree in allocating items. A gold standard generally represents some definite, practically not attainable but only approachable level, determined by a single reference method. There are situations, however, where a gold standard may result from the application of several reference methods or the opinions of several experts, without necessarily achieving a perfect consensus on all items. In a medical context, the

various responses of an expert group may not only reflect the absence of a clear consensus among experienced physicians but also the fuzzy character of the clinical situation at hand. As seen with Williams' syphilis serology data, major discrepancies were observed in the responses given by the 3 reference laboratories for some of the assayed specimens. Therefore, Vanbelle and Albert (2009a) proposed that proficiency testing programs should allow for the fact that a particular non reference laboratory is in perfect agreement with the references laboratories without being in perfect agreement with each of them separately, unlike Schouten's index.

While in theory we may assume that there is always a category of the K-categorical scale with a maximum proportion of raters for each item, it is not necessarily the case in practice. There may indeed be a maximum shared by 2 or more categories, which have to be compared with the category chosen by the isolated rater for this item (see hypothetical laboratory example in Table A.2). However, as mentioned previously, this has virtually no impact on the agreement coefficient obtained. In other terms, two distinct isolated raters will yield the same agreement coefficient (ignoring the weighting scheme) although their response profile is not exactly identical.

In sum, the agreement index proposed by Vanbelle and Albert (2009a) provides a useful alternative to the consensus method and to Light's approach. It also generalizes the agreement index proposed by Schouten (1982) as well as Cohen's kappa coefficient while keeping its attractive properties.

4.10 Proofs

4.10.1 Perfect agreement when $K = 2$

Equivalence 3. *The definition of perfect agreement,* $E(P_i) = E(Y_i) = \pi^{**}$ *and* $corr(P_i, Y_i) = 1$*, is equivalent to writing* $P_i = \pi^{**}(1 - \sqrt{ICC}) + \sqrt{ICC}\ Y_i$.

Proof. Indeed, $\rho = 1$ leads to the linear relation $P_i = a + bY_i$. This implies

$$\begin{aligned} E(P_i) &= \pi^{**} = E(a + bY_i) = a + b\pi^{**} \\ var(P_i) = \sigma^2 &= var(a + bY_i) = b^2 var(Y_i) = b^2\pi^{**}(1 - \pi^{**}). \end{aligned}$$

Thus, $a = (1 - b)\pi^{**}$ and $P_i = (1 - b)\pi^{**} + bY_i$.

Since $ICC = \dfrac{\sigma^2}{\pi(1-\pi)} = b^2 \dfrac{\pi^{**}(1-\pi^{**})}{\pi^{**}(1-\pi^{**})} = b^2$,
we have $P_i = \pi^{**}(1 - \sqrt{ICC}) + \sqrt{ICC}\ Y_i$. ∎

4.10.2 Perfect agreement when $K > 2$

Equivalence 4. *If* Π_M *is defined by*

$$\Pi_M = \sum_{j=1}^{K} E[(\pi_j^{**} + (1 - \pi_j^{**})\sqrt{ICC_j})Y_{ij}]$$

where $E(P_{ij}) = E(Y_{ij}) = \pi_j^{**}$ *and* ICC_j *denotes the intraclass kappa coefficient relative to category* j *(*$j = 1, \cdots, K$*) in the population of raters, we have*

$$\sum_{j=1}^{K} \Pi_{[j]M} = 2\Pi_M + K - 2$$

where $\Pi_{[j]M}$ *corresponds to the quantity described in the binary case (*$K = 2$*) when the nominal scale is dichotomized by grouping all categories other than category* j *together.*

Proof. When the population of raters and the isolated rater are in perfect agreement, we have from Equivalence 1

$$P_{ij} = \pi_j^{**}(1 - \sqrt{ICC_j}) + \sqrt{ICC_j}Y_{ij}.$$

Therefore,

$$\begin{aligned}
\Pi_M = E[\sum_{j=1}^{K} P_{ij}Y_{ij}] &= E[\sum_{j=1}^{K}(\pi_j^{**} + (1 - \pi_j^{**})\sqrt{ICC_j}Y_{ij})Y_{ij}] \\
&= \sum_{j=1}^{K}(\pi_j^{**} + (1 - \pi_j^{**})\sqrt{ICC_j})\pi_j^{**} \\
&= \sum_{j=1}^{K}(\pi_j^{**2} + \sigma_j^{**2}\frac{\sigma_j}{\sigma_j^{**}}) = \sum_{j=1}^{K}(\pi_j^{**2} + \sigma_j\sigma_j^{**}).
\end{aligned}$$

From Equation 4.3,

$$\begin{aligned}
\sum_{j=1}^{K} \Pi_{[j]M} &= \sum_{j=1}^{K}(1 - 2\pi_j^{**}(1 - \pi_j^{**})(1 - \sqrt{ICC_j})) \\
&= \sum_{j=1}^{K}(1 - 2\sigma_j^{**2}\frac{\sigma_j^{**} - \sigma_j}{\sigma_j^{**}}) \\
&= \sum_{j=1}^{K} 1 - 2\sum_{j=1}^{K}\sigma_j^{**2} + 2\sum_{j=1}^{K}\sigma_j^{**}\sigma_j \\
&= K - 2 + 2\sum_{j=1}^{K}(\pi_j^{**2} + \sigma_j\sigma_j^{**}) \\
&= 2\Pi_M + K - 2.
\end{aligned}$$

■

CHAPTER 5

Agreement between two independent groups of raters

5.1 Introduction

Kappa-like agreement indexes to quantify agreement between two raters on a categorical scale were introduced in Chapter 2. They include Cohen's kappa coefficient (Cohen, 1960), the weighted kappa coefficient (Cohen, 1968) and the intraclass kappa coefficient (Kraemer, 1979). All these coefficients are based on the same principle: the proportion of concordant classifications between the two raters (p_o) is corrected for the proportion of concordant classifications expected by chance (p_e) and standardized $\hat{\kappa} = (p_o - p_e)/(1 - p_e)$ to obtain a value 1 when agreement between the two raters is perfect and 0 in case of agreement due to chance alone. Although agreement is often searched between two individual raters, there are situations where agreement is needed between two groups of raters. For example, a group of students may be evaluated against another group of students or against a group of experts, each group classifying the same set of items on a categorical scale. Likewise, agreement may be searched between two groups of physicians with different specialties or professional experience in diagnosing patients by means of the same (positive/negative) clinical test. In such instances, each group is seen as a whole, a global entity with its own heterogeneity. Interest resides in the overall degree of agreement between the groups, not in the agreement between individuals themselves. In fact, the groups may perfectly agree while some of their members may not.

Methods testing for evidence of agreement between two groups of raters when ordering items were proposed by Schucany and Frawley (1973), Hollander and Sethuraman (1978), Kraemer (1981) and Feigin and Alvo (1986). These methods are generally based on the Spearman rank correlation coefficient or Kendall's tau coefficient. However, methods designed to quantify the degree of agreement between two groups of raters on a nominal or ordinal scale barely exist and it appears that the only reference found in the literature is a paper written by Schouten (1982). He developed a measure of pairwise interobserver agreement between two groups of raters to find clusters of homogeneous subgroups of raters when all raters classify the items on a categorical scale. His method consists in substituting in the kappa coefficient the observed proportion of agreement (p_o) and the proportion of agreement expected by chance (p_e) by, respectively, the mean of the observed ($\overline{p}_o$) and of the expected ($\overline{p}_e$) proportions of agreement obtained between all possible pairs of raters formed with one rater in each group, namely $\hat{\kappa} = (\overline{p}_o - \overline{p}_e)/(1 - \overline{p}_e)$. Unfortunately, in Schouten's approach, perfect agreement between the two groups can only be achieved if there is perfect agreement within each group.

Although there is a clear lack of theoretical work on agreement measures between two groups of raters, it is common practice in the applied literature to determine empirically a consensus category in each group of raters in order to reduce the problem to the case of two raters. To our knowledge, the consensus method is used as an intuitive method and there is no theoretical proof to justify its use. The consensus category may be defined as the modal category (e.g., van Hoeij et al. (2004)), the median category (e.g., Raine et al. (2004)) or the mean category (e.g., Bland et al. (2005)) if the scale is ordinal. When a consensus category is found in each group for each item, the agreement between these categories is studied in the usual way (case of two raters). In all instances, however, the question of how to proceed when a consensus can not be reached remains. Moreover, different rules to define the consensus category may lead to different conclusions (Kraemer et al., 2004). Indeed, consider a group of 10 raters allocating an item on a 5-point Likert scale and suppose that 3 raters classify the item in category 1, 2 in category 2, none in categories 3 and 4, and 5 in category 5. The consensus category defined by the modal rule is category 5, by the median rule category 2, 3, 4 or 5 and by the mean rule category 3 (category chosen by none of the raters in the group). The three rules may almost inevitably lead to three different conclusions. It should also be remarked that consensus does not take into account the variability in the groups in the sense that different patterns of responses may lead to the determination of the same consensus category and thus lead to the same conclusions. Indeed, in the example above, if 6 instead of 5 raters classified the item in category 5, the modal category would still be category 5, leading to the

same conclusion although the variability in the group is different.

The present chapter aimed at defining an overall agreement index between two groups of raters, taking into account the heterogeneity of each group. Furthermore, the agreement index overcomes the problem of consensus and can be viewed as a natural extension of Cohen's kappa coefficient to two groups of raters. The novel agreement index was defined on a population-based model (Vanbelle and Albert, 2009b) and its sampling variability determined by the Jackknife method (Efron and Tibshirani, 1993).

5.2 The two group agreement index

5.2.1 Binary scale

Consider a population of items $\mathcal{I}$ and two distinct populations of raters $\mathcal{R}_1$ and $\mathcal{R}_2$. Suppose that items have to be classified in two categories ($K = 2$). Now, consider a randomly selected rater r from population $\mathcal{R}_g$ and a randomly selected item i from population $\mathcal{I}$. Let $X_{ir,g}$ be the random variable such that $X_{ir,g} = 1$ if rater r of population $\mathcal{R}_g$ classifies item i in category 1 and $X_{ir,g} = 0$ otherwise. For each item i, $E(X_{ir,g}|i) = P(X_{ir,g} = 1) = P_{i,g}$ over the population of raters. Then, over the population of items, $E(P_{i,g}) = E[E(X_{ir,g}|i)] = \pi_g$ and $var(P_{i,g}) = \sigma_g^2$. Finally, let

$$ICC_g = \frac{\sigma_g^2}{\pi_g(1-\pi_g)}$$

be the intraclass correlation coefficient in group g ($g = 1, 2$) denoted by ICC_g for convenience (see Chapter 2, Section 2.4.1). The joint distribution of the classifications of item i made by the two populations of raters consists of four probabilities summing up to 1, $(1 - P_{i,1})(1 - P_{i,2})$, $(1 - P_{i,1})P_{i,2}$, $P_{i,1}(1 - P_{i,2})$ and $P_{i,1}P_{i,2}$. For example, $P_{i,1}P_{i,2}$ denotes the probability that both populations of raters classify item i into category 1. The expectations of these joint probabilities over the population of items $\mathcal{I}$ can be represented in a 2×2 classification table, as displayed in Table 5.1 with $\rho = corr(P_{i,1}, P_{i,2}) = [E(P_{i,1}P_{i,2}) - \pi_1\pi_2)]/\sigma_1\sigma_2$, the correlation over $\mathcal{I}$ between the random variables $P_{i,1}$ and $P_{i,2}$.

The probability that the two populations of raters agree on the classification of item i is naturally defined by

$$\Pi_i = P_{i,1}P_{i,2} + (1 - P_{i,1})(1 - P_{i,2}). \tag{5.1}$$

Thus, at the population level, the *mean probability of agreement* over $\mathcal{I}$ is (see Table 5.1)

$$\Pi_T = E(\Pi_i) = \pi_1\pi_2 + (1 - \pi_1)(1 - \pi_2) + 2\rho\sigma_1\sigma_2. \tag{5.2}$$

Table 5.1. Expected joint classification probabilities of the two populations of raters over the population of items

		$\mathcal{R}_2$		
		0	1	
$\mathcal{R}_1$	0	$E[(1-P_{i,1})(1-P_{i,2})]$ $(1-\pi_1)(1-\pi_2)+\rho\sigma_1\sigma_2$	$E[(1-P_{i,1})P_{i,2}]$ $(1-\pi_1)\pi_2-\rho\sigma_1\sigma_2$	$1-\pi_1$
	1	$E[P_{i,1}(1-P_{i,2})]$ $\pi_1(1-\pi_2)-\rho\sigma_1\sigma_2$	$E[P_{i,1}P_{i,2}]$ $\pi_1\pi_2+\rho\sigma_1\sigma_2$	π_1
		$1-\pi_2$	π_2	1

This quantity does not only involve the marginal probabilities that populations $\mathcal{R}_1$ and $\mathcal{R}_2$ classify items in category 1 (π_1 and π_2) but also the variability within each population of raters (σ_1 and σ_2) and the correlation ρ.

Under the assumption of random assignment of item i by the two populations of raters ($E[P_{i,1}P_{i,2}] = E[P_{i,1}]E[P_{i,2}]$), the *mean probability of agreement expected by chance* is simply the product of the marginal probabilities, namely

$$\Pi_E = \pi_1\pi_2 + (1-\pi_1)(1-\pi_2). \tag{5.3}$$

It is seen that this quantity can be obtained by setting the correlation coefficient ρ equal to 0 in Equation 5.2, or equivalently by setting either σ_1^2 and/or σ_2^2 equal to 0.

Vanbelle and Albert (2009b) defined the agreement index between the two populations of raters in a kappa-like way, namely

$$\kappa = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E} \tag{5.4}$$

where $\Pi_M = max(\Pi_T)$ corresponds to the *maximum attainable value of the mean probability of agreement* (Equation 5.2) given the existing heterogeneity in each population of raters. Thus, $\kappa = 1$ when agreement is perfect, $\kappa = 0$ when agreement is only due to chance and $\kappa < 0$ when agreement is less than one would expect by chance.

There is a need at this stage of the development to explicit the notion of "perfect agreement" ($\kappa = 1$). By definition, the two populations of raters are said to be in perfect agreement if and only if $P_{i,1} = P_{i,2} = P_i$, for all items i in $\mathcal{I}$ (Vanbelle and Albert, 2009b). In other words, the two populations of raters "perfectly" agree if and only if the probability of classifying an item in a given category is the same for the two populations. Intuitively, it is obvious that if the probability of classifying

item i in category 1 is different in the two populations of raters, the latter can not agree perfectly. Note that the present definition extends that of perfect agreement between two raters, namely that $X_{i,1} = X_{i,2} = X_i$ for each item i. Under the definition of perfect agreement, if we write $E(P_i) = \pi$ and $var(P_i) = \sigma^2$, we have $ICC_g = ICC = \sigma^2/\pi(1-\pi)$, $(g = 1, 2)$ and Π_M is then given by the expression

$$\Pi_M = E(\Pi_i) = 2\sigma^2 + 2\pi^2 - 2\pi + 1 = 1 - 2\pi(1-\pi)(1-ICC). \tag{5.5}$$

It is seen that $\Pi_M = 1$ if the intraclass kappa coefficient is equal to 1 in both populations of raters ($ICC = 1$, i.e. perfect agreement within each population), and/or trivially if $\pi = 0$ or $\pi = 1$ (no variability in the allocation process). Note that Schouten's agreement index is given by Equation 5.4 where $\Pi_M = 1$.

An intraclass version of κ can be derived using the additional assumption $\pi_1 = \pi_2 = \pi$ (equality of marginal probabilities). In that case, we have

$$\kappa_I = \frac{E(P_{i,1}P_{i,2}) - \pi^2}{\sigma^2} \tag{5.6}$$

which is equivalent to the correlation coefficient between $P_{i,1}$ and $P_{i,2}$ under the assumption of equal marginal probabilities.

5.2.2 Nominal scale

When $K > 2$, Vanbelle and Albert (2009b) defined the coefficient of agreement between two independent populations of raters by

$$\kappa = \frac{\sum_{j=1}^{K}(\Pi_{[j]T} - \Pi_{[j]E})}{\sum_{j=1}^{K}(\Pi_{[j]M} - \Pi_{[j]E})} = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E} \tag{5.7}$$

where the quantities $\Pi_{[j]T}$, $\Pi_{[j]E}$ and $\Pi_{[j]M}$ correspond to the quantities described in the dichotomous case when the nominal scale is dichotomized by grouping all categories other than category j together and Π_T, Π_E and Π_M are defined by

$$\Pi_T = \sum_{j=1}^{K} E(P_{ij,1}P_{ij,2});\ \Pi_E = \sum_{j=1}^{K} \pi_{j,1}\pi_{j,2};\ \text{and } \Pi_M = \sum_{j=1}^{K} E(P_{ij}^2)$$

and extend naturally the quantities defined in the dichotomous case. Indeed, $P_{ij,g}$ denotes the probability for item i to be classified in category j $(j = 1, \cdots, K)$ by the population of raters $\mathcal{R}_g$ $(g = 1, 2)$ and is a random variable over the population of items $\mathcal{I}$. We have $P_{ij,g} = P(X_{ijr,g} = 1|i)$ where the binary random variable $X_{ijr,g}$ is equal to 1 if rater r of population $\mathcal{R}_g$ classifies item i in category j and $\sum_{j=1}^{K} P_{ij,g} = 1$. Over the population of items $\mathcal{I}$, $E(P_{ij,g}) = \pi_{j,g}$ $(g = 1, 2)$. The equivalence of the two expressions in Equation 5.7 is proven in Section 5.9. The two populations of raters are defined to be in perfect agreement if and only if $P_{ij,1} = P_{ij,2} = P_{ij}$ for all items i in $\mathcal{I}$ $(j = 1, \cdots, K)$, extending the definition of the dichotomous case.

5.2.3 Ordinal scale

A weighted version of the agreement index between two populations of raters, accounting for the fact that some disagreements may be more important than others, is defined in the same way as the weighted kappa coefficient (Cohen, 1968). We have

$$\kappa_W = \frac{\Pi_{T,W} - \Pi_{E,W}}{\Pi_{M,W} - \Pi_{E,W}} \tag{5.8}$$

where

$$\Pi_{T,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E(P_{ij,1} P_{ik,2}), \tag{5.9}$$

$$\Pi_{E,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} \pi_{j,1} \pi_{k,2}, \tag{5.10}$$

$$\Pi_{M,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E(P_{ij} P_{ik}). \tag{5.11}$$

The unweighted agreement index κ (see Equation 5.7) is obtained by using the weighting scheme $w_{jk} = 1$ if $j = k$ and $w_{jk} = 0$ otherwise ($j \neq k \in 1, \cdots, K$).

5.3 Estimation of the parameters

Consider a random sample of N items from $\mathcal{I}$, a random sample of R_1 raters from $\mathcal{R}_1$ (group G_1) and a random sample of R_2 raters from $\mathcal{R}_2$ (group G_2).

5.3.1 Binary scale

Suppose that $x_{ir,g}$ denote the observed values of the random variables $X_{ir,g}$ defined in Section 5.2.1 ($i = 1, \cdots, N; r = 1, \cdots, R_g, ; g = 1, 2$). Let

$$n_{i,g} = \sum_{r=1}^{R_g} x_{ir,g}$$

denote the number of raters of group G_g classifying item i in category 1 ($g = 1, 2$). Then, let

$$p_{i,g} = \frac{n_{i,g}}{R_g}$$

be the corresponding proportions ($i = 1, \cdots, N; j = 1, \cdots, K; g = 1, 2$).

At the population level, the *mean agreement over the population of items* $\mathcal{I}$ between the two populations of raters, Π_T, is estimated by the *observed proportion of agreement*

$$\widehat{\Pi}_T = p_o = \frac{1}{N}\sum_{i=1}^{N}[p_{i,1}p_{i,2} + (1-p_{i,1})(1-p_{i,2})]. \tag{5.12}$$

Likewise, the *mean probability of agreement expected by chance*, Π_E, is estimated by the *proportion of agreement expected by chance*

$$\widehat{\Pi}_E = p_e = p_1p_2 + (1-p_1)(1-p_2) \tag{5.13}$$

where $p_g = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} p_{i,g}$ $(g = 1, 2)$.
The agreement index between the two populations of raters is then estimated by

$$\widehat{\kappa} = \frac{p_o - p_e}{p_m - p_e} \tag{5.14}$$

where p_m corresponds to the maximum possible proportion of agreement derived from the samples. Indeed, recall that Π_M is obtained when $P_{i,1} = P_{i,2} = P_i$ and corresponds to the maximum expected agreement over the population of items. Thus, given the observed data, the maximum observed proportion of agreement can be obtained when $p_i = p_{i,g}$ $(g = 1, 2)$, leading to $p_o = p_{i,g}^2 + (1-p_{i,g})^2$. Since $p_{i,1}p_{i,2} + (1-p_{i,1})(1-p_{i,2}) \leq max_g[p_{i,g}^2 + (1-p_{i,g})^2]$ for each item i, it follows that

$$\widehat{\Pi}_M = p_m = \frac{1}{N}\sum_{i=1}^{N} max_g[p_{i,g}^2 + (1-p_{i,g})^2]. \tag{5.15}$$

It is seen that if $p_{i,1} = p_{i,2}$ $(i = 1, \cdots, N)$, $p_o = p_m$ and $\hat{\kappa} = 1$.

5.3.2 Nominal scale

Let $x_{ijr,g}$ denote the observed values of the random variables $X_{ijr,g}$ equal to 1 if rater r $(r = 1, \cdots, R_g)$ of population $\mathcal{R}_g$ $(g = 1, 2)$ classifies item i $(i = 1, \cdots, N)$ in category j $(j = 1, \cdots, K)$. The assessment of the N items by the two groups of raters can be conveniently summarized in a two-way classification table as seen in Table 5.2. Let

$$n_{ij,g} = \sum_{r=1}^{R_g} x_{ijr,g}$$

denote the number of raters of group G_g classifying item i in category j $(g = 1, 2)$. Then, let

$$p_{ij,g} = \frac{n_{ij,g}}{R_g}$$

be the corresponding proportions ($i = 1, \cdots, N; j = 1, \cdots, K; g = 1, 2$). We have $\sum_{j=1}^{K} p_{ij,g} = 1$, ($i = 1, \cdots, N; g = 1, 2$). Finally, let

$$c_{jk} = \frac{1}{N} \sum_{i=1}^{N} p_{ij,1} p_{ik,2} \ (j, k = 1, \cdots, K).$$

The quantities c_{jk} estimate the joint probability that populations $\mathcal{R}_1$ and $\mathcal{R}_2$ classify a randomly selected item i in category j and k, respectively ($c_{jk} = E(\widehat{P_{ij,1} P_{ik,2}})$; $j, k = 1, \cdots, K$). A $K \times K$ matrix can then be derived from the original data (see Table 5.2).

Table 5.2. Two-way classification table of the N items by the two groups of raters on a K-categorical scale

		G_2					
	Category	1	...	j	...	K	Total
	1	c_{11}	...	c_{1j}	...	c_{1K}	$c_{1.}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_1	j	c_{j1}	...	c_{jj}	...	c_{jK}	$c_{j.}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	K	c_{K1}	...	c_{Kj}	...	c_{KK}	$c_{K.}$
	Total	$c_{.1}$	...	$c_{.j}$	...	$c_{.K}$	1

The *mean probability of agreement* between the two populations of raters, Π_T, is estimated by

$$\widehat{\Pi}_T = p_o = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij,1} p_{ij,2} = \sum_{j=1}^{K} c_{jj} \tag{5.16}$$

and the *mean probability of agreement expected by chance*, Π_E, is estimated by

$$\widehat{\Pi}_E = p_e = \sum_{j=1}^{K} p_{j,1} p_{j,2} = \sum_{j=1}^{K} c_{j.} c_{.j} \tag{5.17}$$

where $p_{j,g} = \frac{1}{N} \sum_{i=1}^{N} p_{ij,g}$.

The agreement index between the two populations of raters is then estimated as before by

$$\widehat{\kappa} = \frac{p_o - p_e}{p_m - p_e} \tag{5.18}$$

where

$$p_m = \frac{1}{N} \sum_{i=1}^{N} max(\sum_{j=1}^{K} p_{ij,1}^2, \sum_{j=1}^{K} p_{ij,2}^2) \tag{5.19}$$

is the maximum possible proportion of agreement derived from the data, obtained by extending the argument developed for the dichotomous case. Note that when there is only one rater in each group of raters ($R_1 = R_2 = 1$), the agreement coefficient $\hat{\kappa}$ merely reduces to Cohen's κ coefficient (Cohen, 1960).

5.3.3 Ordinal scale

The weighted version of the agreement index is estimated in exactly the same way, namely

$$\widehat{\kappa}_W = \frac{p_{o,W} - p_{e,W}}{p_{m,W} - p_{e,W}} \tag{5.20}$$

with

$$\widehat{\Pi}_{T,W} = p_{o,w} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{ij,1} p_{ik,2} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} c_{jk}, \tag{5.21}$$

$$\widehat{\Pi}_{E,W} = p_{e,w} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{j,1} p_{k,2} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} c_{j.} c_{.k} \tag{5.22}$$

and

$$\widehat{\Pi}_{M,W} = p_{m,W} = \frac{1}{N} \sum_{i=1}^{N} max(\sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{ij,1} p_{ik,1}, \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{ij,2} p_{ik,2}). \tag{5.23}$$

5.3.4 Sampling variability

The Jackknife method (Efron and Tibshirani, 1993) can be used to determine the sampling variance of the agreement indexes, as explained in Section 4.3.4.

5.3.5 Example

Consider the following hypothetic example to illustrate how to compute the proposed agreement index. Suppose that a group G_1 of 12 raters and a group G_2 of 3 raters have to classify 3 items on a 5-point Likert scale ranging from (-2) to (2) (Table 5.3).

The responses given by the 2 groups of raters are then summarized in Table 5.4 and expressed in terms of proportions ($p_{ij,g} = n_{ij,g}/R_g$) ($g = 1, 2$) in Table 5.5. The marginal classification distributions of the groups of raters ($p_{j,g}$) are also determined.

Table 5.3. Responses given by the groups G_1 ($R_1 = 12$) and G_2 ($R_2 = 3$) for 3 items on a 5-point Likert scale (hypothetic example)

	Group G_1												Group G_2		
Item	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3
1	0	1	2	2	2	1	2	1	1	1	1	1	1	2	1
2	0	-1	1	0	0	-1	-1	0	0	-1	-1	-1	0	2	2
3	1	1	-2	-1	-1	1	-2	-2	-1	-1	1	1	-2	-1	-2

Table 5.4. Summary of the responses given by the groups of raters G_1 and G_2 (hypothetic example)

	Group G_1					Group G_2				
	Category					Category				
Item	(-2)	(-1)	(0)	(1)	(2)	(-2)	(-1)	(0)	(1)	(2)
1	0	0	1	7	4	0	0	0	2	1
2	0	6	5	1	0	0	0	1	0	2
3	3	4	0	5	0	2	1	0	0	0

Table 5.5. Distribution of the responses given by the two groups of raters G_1 and G_2 (hypothetic example)

	Group G_1					Group G_2				
	Category					Category				
Item	(-2)	(-1)	(0)	(1)	(2)	(-2)	(-1)	(0)	(1)	(2)
1	0	0	0.08	0.58	0.33	0	0	0	0.66	0.33
2	0	0.50	0.42	0.08	0	0	0	0.33	0	0.66
3	0.25	0.33	0	0.42	0	0.66	0.33	0	0	0
p_j	0.08	0.28	0.17	0.36	0.11	0.22	0.11	0.11	0.22	0.33

The observed proportion of agreement is equal to

$$\begin{aligned} p_o &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij,1}p_{ij,2} \\ &= (0.58 \times 0.66 + 0.33 \times 0.33 + 0.42 \times 0.33 + 0.25 \times 0.66 + 0.33 \times 0.33)/3 \\ &= 0.31. \end{aligned}$$

The proportion of agreement expected by chance is equal to

$$\begin{aligned} p_e &= \sum_{j=1}^{K} p_{j,1}p_{j,2} \\ &= 0.08 \times 0.22 + 0.28 \times 0.11 + 0.17 \times 0.11 + 0.36 \times 0.22 + 0.11 \times 0.33 = 0.19. \end{aligned}$$

To determine the maximum possible observed proportion of agreement, each group is duplicated to artificially have perfect agreement and the observed proportion of agreement is calculated (see Table 5.6).

Table 5.6. Squared proportion of raters classifying each item in the 5 categories as explained in Equation 5.15 (hypothetic example)

	Group G_1						Group G_2					
	Category					Sum	Category					Sum
Item	(-2)	(-1)	(0)	(1)	(2)		(-2)	(-1)	(0)	(1)	(2)	
1	0	0	0.01	0.34	0.11	0.46	0	0	0	0.44	0.11	0.56
2	0	0.25	0.17	0.01	0	0.43	0	0	0.11	0	0.44	0.56
3	0.06	0.11	0	0.17	0	0.35	0.44	0.11	0	0	0	0.56

For each item, the highest observed proportion of agreement is chosen. It leads to the maximum proportion of observed agreement

$$p_m = \frac{0.56 + 0.56 + 0.56}{3} = 0.56.$$

The agreement index between the two groups is then equal to

$$\hat{\kappa} = \frac{p_o - p_e}{p_m - p_e} = \frac{0.31 - 0.19}{0.56 - 0.19} = 0.33.$$

5.4 Consensus approach

5.4.1 Binary scale

Consider a population of items $\mathcal{I}$ and two distinct populations of raters $\mathcal{R}_1$ and $\mathcal{R}_2$. Suppose that items have to be classified in two categories ($K = 2$). Let $\mathcal{I}_C$

denote the sub-population of items on which a consensus (C) is always possible in both populations of raters. In $\mathcal{I}_C$, consider the random variable $Z_{i,g}$ such that $Z_{i,g} = 1$ if there is a consensus on category 1 for item i in the population $\mathcal{R}_g$ and $Z_{i,g} = 0$ otherwise. The agreement index based on the consensus method then reduces to the case of two raters, the consensus defining a single rater in each group. Then, over $\mathcal{I}_C$, let $E(Z_{i,g}) = \phi_g$ and $var(Z_{i,g}) = \sigma_g'^2 = \phi_g(1-\phi_g)$. If ρ' denotes the correlation coefficient between $Z_{i,1}$ and $Z_{i,2}$, we have the following representation of the expected probabilities between the two consensus values (Table 5.7).

Table 5.7. Expected probabilities of the classification of the two populations of raters over the sub-population $\mathcal{I}_C$ of items where a consensus exists

		$\mathcal{R}_2$		
		0	1	
$\mathcal{R}_1$	0	$E[(1-Z_{i,1})(1-Z_{i,2})]$ $(1-\phi_1)(1-\phi_2)+\rho'\sigma_1'\sigma_2'$	$E[(1-Z_{i,1})Z_{i,2}]$ $(1-\phi_1)\phi_2-\rho'\sigma_1'\sigma_2'$	$1-\phi_1$
	1	$E[Z_{i,1}(1-Z_{i,2})]$ $\phi_1(1-\phi_2)-\rho'\sigma_1'\sigma_2'$	$E(Z_{i,1}Z_{i,2})$ $\phi_1\phi_2+\rho'\sigma_1'\sigma_2'$	ϕ_1
		$1-\phi_2$	ϕ_2	1

The agreement between the two populations of raters on item i based on the consensus, denoted Π_{iC}, is defined by

$$\Pi_{iC} = Z_{i,1}Z_{i,2} + (1-Z_{i,1})(1-Z_{i,2}). \tag{5.24}$$

Thus,

$$E(\Pi_{iC}) = \Pi_{TC} = \phi_1\phi_2 + (1-\phi_1)(1-\phi_2) + 2\rho'\sigma_1'\sigma_2'. \tag{5.25}$$

The agreement expected by chance is defined by

$$\Pi_{EC} = \phi_1\phi_2 + (1-\phi_1)(1-\phi_2) \tag{5.26}$$

and perfect agreement is achieved when $Z_{i,1} = Z_{i,2}$, for all items in $\mathcal{I}_C$, leading to

$$E(\Pi_{iC}) = \Pi_{MC} = 1.$$

Therefore, the agreement coefficient between the two populations of raters is defined by

$$\kappa_C = \frac{\Pi_{TC} - \Pi_{EC}}{1 - \Pi_{EC}}. \tag{5.27}$$

5.4.2 Nominal scale

Consider the random variable $Z_{ij,g}$ such that $Z_{ij,g} = 1$ if there is a consensus on category j for item i in population $\mathcal{R}_g$ and $Z_{ij,g} = 0$ otherwise. Then, over $\mathcal{I}_C$, let $E(Z_{ij,g}) = \phi_{j,g}$. In the same way as before,

$$\kappa_C = \frac{\sum_{j=1}^{K}(\Pi_{[j]TC} - \Pi_{[j]EC})}{\sum_{j=1}^{K}(\Pi_{[j]MC} - \Pi_{[j]EC})} = \frac{\Pi_{TC} - \Pi_{EC}}{\Pi_{MC} - \Pi_{EC}} \tag{5.28}$$

where $\Pi_{[j]TC}$, $\Pi_{[j]EC}$ and $\Pi_{[j]MC}$ correspond to the quantities described in the dichotomous case when the nominal scale is dichotomized by grouping all categories other than category j together. The quantities Π_{TC}, Π_{EC} and Π_{MC} are defined respectively by

$$\Pi_{TC} = \sum_{j=1}^{K} E(Z_{ij,1}Z_{ij,2});\ \Pi_{EC} = \sum_{j=1}^{K} \phi_{j,1}\phi_{j,2};\ \Pi_{MC} = 1.$$

5.4.3 Ordinal scale

The weighted version of the consensus approach can also be derived in the same way as before by introducing weights in the expression of Π_{TC}, Π_{EC} and Π_{MC}.

$$\Pi_{T,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}E(Z_{ij,1}Z_{ik,2}); \tag{5.29}$$

$$\Pi_{E,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}\phi_{j,1}\phi_{k,2}; \tag{5.30}$$

$$\Pi_{M,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}E(Z_{ij}Z_{ik}) = 1 \tag{5.31}$$

leading to

$$\kappa_{C,W} = \frac{\Pi_{T,WC} - \Pi_{E,WC}}{1 - \Pi_{E,WC}}. \tag{5.32}$$

5.4.4 Estimation of the parameters

Consider again a random sample of R_1 raters from $\mathcal{R}_1$, a random sample of R_2 raters from $\mathcal{R}_2$ and a random sample of N items from $\mathcal{I}$. Let N_C ($\leq N$) denote the number of items where a consensus exist in each group. Suppose that $z_{ij,g}$ denotes the observed values of the random variables $Z_{ij,g}$ ($i = 1, \cdots, N_C; j = 1, \cdots, K; g = 1, 2$) defined in the previous section. The assessment of the N_C items on which the

two groups of raters can determine a consensus can be conveniently summarized by

$$d_{jk} = \frac{1}{N_C} \sum_{i=1}^{N_C} z_{ij,1} z_{ik,2} \ (j,k = 1, \cdots, K).$$

Similarly to what was done in Section 5.3, the *observed weighted agreement* between the two groups of raters is obtained by

$$\widehat{\Pi}_{T,WC} = p_{o,WC} = \frac{1}{N_C} \sum_{i=1}^{N_C} \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} z_{ij,1} z_{ik,2} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} d_{jk} \tag{5.33}$$

and the *agreement expected by chance* by the expression

$$\widehat{\Pi}_{E,WC} = p_{e,WC} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} z_{j,1} z_{k,2} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} d_{j.} d_{.k} \tag{5.34}$$

where $z_{j,g} = \dfrac{1}{N_C} \displaystyle\sum_{i=1}^{N_C} z_{ij,g}$, $(g = 1, 2)$ leading to the weighted agreement coefficient

$$\hat{\kappa}_{C,W} = \frac{p_{o,WC} - p_{e,WC}}{1 - p_{e,WC}}. \tag{5.35}$$

5.4.5 Example

Consider the example developed in Section 5.3.5 to illustrate the consensus method. A consensus is determined in each group using the majority rule i.e., the consensus category is determined for each item as the category the most chosen by the raters in the group. The data are then summarized in a 5×5 contingency table by cross-classifying the consensuses found in the two groups of raters (see Table 5.8).

Table 5.8. Cross-classification of the responses given by the two group of raters (consensus)

		Group G_2					
	Category	(-2)	(-1)	(0)	(1)	(2)	Total
Group G_1	(-2)	0	0	0	0	0	0
	(-1)	0	0	0	0	0.33	0.33
	(0)	0	0	0	0	0	0
	(1)	0.33	0	0	0.33	0	0.66
	(2)	0	0	0	0	0	0
	Total	0.33	0	0	0.33	0.33	1

The observed proportion of agreement is equal to

$$p_{oC} = \sum_{j=1}^{K} d_{jj} = 0.33,$$

and the proportion of agreement expected by chance to

$$p_{eC} = \sum_{j=1}^{K} d_{j.} d_{.j} = 0 \times 0 + 0 \times 0.33 + 0.33 \times 0 + 0.33 \times 0.66 + 0.33 \times 0 = 0.22.$$

This leads a Cohen's kappa coefficient of

$$\widehat{\kappa}_C = \frac{p_{oC} - p_{eC}}{1 - p_{eC}} = \frac{0.33 - 0.22}{1 - 0.22} = 0.14.$$

5.5 Schouten's agreement index

Schouten (1982) also developed a hierarchical clustering method, consisting in grouping the raters with the highest inter-cluster agreement coefficient.

5.5.1 Definition

For two clusters G_1 and G_2 consisting of R_1 and R_2 raters, where no rater belongs to G_1 and G_2 simultaneously, Schouten (1982) defined the inter-cluster kappa coefficient by

$$\hat{\kappa}_{G_1,G_2}(w) = \frac{o_{G_1,G_2}(w) - e_{G_1,G_2}(w)}{1 - e_{G_1,G_2}(w)} \tag{5.36}$$

where

$$o_{G_1,G_2}(w) = \frac{1}{R_1 R_2} \sum_{r \in G_1} \sum_{s \in G_2} o_{r,s}(w) \tag{5.37}$$

and

$$e_{G_1,G_2}(w) = \frac{1}{R_1 R_2} \sum_{r \in G_1} \sum_{s \in G_2} e_{r,s}(w). \tag{5.38}$$

The quantities $o_{r,s}(w)$ and $e_{r,s}(w)$ were defined in Chapter 4 (Section 4.5).

5.5.2 Hierarchical clustering

When the number of raters is large, Schouten (1982) proposed to divide the group of raters into several homogeneous subgroups, with higher degree of pairwise inter-rater agreement within subgroups than between subgroups, and to find out why and in which way different subgroups differ in opinion. Schouten (1982) used the weighted kappa coefficient defined in Equation 5.36 to identify such homogeneous subgroups, called "clusters".

The hierarchical cluster analysis starts with R_1+R_2 clusters formed by the R_1+R_2 raters of the group. Next, the raters within the two clusters with the highest inter-cluster kappa coefficient are grouped together and form a new cluster, and this may go on until finally all raters are considered to be in one cluster.

5.5.3 Example

Consider the hypothetical example described in Table 5.3. The observed proportion of agreement and the proportion of agreement expected by chance corresponding to each pair formed by one rater in the group G_1 and one in the group G_2 are given in Table 5.9.

Table 5.9. Observed proportion of agreement (p_o), expected proportion of agreement (p_e) and Cohen's kappa coefficients ($\hat{\kappa}$) between each rater of the group G_1 and each rater of the group G_2

		p_o			p_e			$\hat{\kappa}$		
		Group G_2			Group G_2			Group G_2		
	Rater	1	2	3	1	2	3	1	2	3
Group G_1	1	0.33	0	0	0.33	0	0.11	0	0	-0.13
	2	0.33	0	0.33	0.22	0.11	0.22	0.14	-0.13	0.14
	3	0.33	0.33	0.33	0.22	0.22	0.33	0.14	0.14	0
	4	0.33	0.67	0	0.11	0.33	0.11	0.25	0.50	-0.13
	5	0.33	0.67	0	0.11	0.33	0.11	0.25	0.50	-0.13
	6	0.33	0	0.33	0.22	0.11	0.22	0.14	-0.13	0.14
	7	0.33	0.33	0.33	0.11	0.33	0.22	0.25	0	0.14
	8	1	0	0.67	0.33	0	0.22	1	0	0.57
	9	0.67	0.33	0.33	0.22	0.11	0.11	0.57	0.25	0.25
	10	0.33	0.33	0.33	0.11	0.22	0.11	0.25	0.14	0.25
	11	0.33	0	0.33	0.22	0.11	0.22	0.14	-0.13	0.14
	12	0.33	0	0.33	0.22	0.11	0.22	0.14	-0.13	0.14
	Mean	0.42	0.22	0.28	0.20	0.17	0.19	0.27	0.09	0.12

This leads to a Schouten's agreement index of

$$\begin{aligned} \hat{\kappa}_{G_1,G_2} &= \frac{o_{G_1,G_2} - e_{G_1,G_2}}{1 - e_{G_1,G_2}} \\ &= \frac{(0.42 + 0.22 + 0.28)/3 - (0.20 + 0.17 + 0.19)/3}{1 - (0.20 + 0.17 + 0.19)/3} \\ &= \frac{0.31 - 0.19}{1 - 0.19} = 0.15. \end{aligned} \tag{5.39}$$

Remark that the mean of the kappa coefficients (=0.16) is near Schouten's index (=0.15) but not equal.

5.6 Comparison of the agreement indexes

5.6.1 With the consensus method

The consensus approach is equivalent to the new agreement index if and only if $R_1 = R_2 = 1$ or if and only if a consensus is always possible for each item in both populations of raters ($\mathcal{I}_C = \mathcal{I}$) and there is perfect agreement in both populations of raters ($P_{ij,1} = P_{ij,2} = P_{ij}$, $\forall i$).

5.6.2 With Schouten's index

As in the previous chapter, we can easily show that $p_{o,w}$ and $o_{G_1,G_2}(w)$ are equivalent as well as $p_{e,w}$ and $e_{G_1,G_2}(w)$. With the additional assumption $ICC_1 = ICC_2 = 1$, i.e., there is perfect agreement in each population of raters, the proposed agreement index κ is algebraically equivalent to the inter-cluster agreement index introduced by Schouten (1982).

5.7 Script Concordance Test

Let look again to the example of the SCT developed in Chapter 4, Section 4.8.2 and consider now the 39 students training in "general practice" as a whole group. We thus have in the present example $R_1 = 11$, $R_2 = 39$, $N = 34$ and $K = 5$. The cross-classification matrix $(c_{jk}, j, k = 1, \cdots, 5)$ between the group of medical students and the group of experts is given in Table 5.10.

Table 5.10. Two-way classification table of the 34 items of the Script Concordance Test (SCT) by the group of 11 medical experts and by the group of 39 medical students using a 5-point Likert scale ((-2) The assumption is practically eliminated; (-1) The assumption becomes less likely; (0) The information has no effect on the assumption; (+1) The assumption becomes more likely (+2) The assumption is practically the only possible)

		Medical experts					
		(-2)	(-1)	(0)	(1)	(2)	Total
Medical students	(-2)	0.077	0.054	0.028	0.009	0.002	0.170
	(-1)	0.036	0.067	0.066	0.033	0.012	0.214
	(0)	0.022	0.053	0.187	0.062	0.013	0.337
	(1)	0.013	0.026	0.069	0.090	0.025	0.223
	(2)	0.005	0.009	0.013	0.020	0.010	0.057
	Total	0.153	0.209	0.363	0.214	0.057	1

Since the scale is ordinal, weighted agreement indexes were calculated using the

quadratic weighting scheme ($w_{jk} = 1 - (|k-j|/4)^2$, $k, j = -2, \cdots, 2$) (Fleiss and Cohen, 1973). On the basis of the study material, we found that the observed proportion of agreement, the proportion of agreement expected by chance and the maximum proportion of agreement were respectively $p_{o,w} = 0.80$, $p_{e,w} = 0.69$ and $p_{m,w} = 0.84$, yielding a weighted agreement index $\hat{\kappa}_W = (0.80 - 0.69)/(0.84 - 0.69) = 0.72$.

Table 5.11. Weighted agreement indexes between the group of 11 experts and the group of 39 students for the Script Concordance Test (SCT) with 34 items obtained by four different methods with quadratic weighting scheme.

Method	Coefficient	N	p_o	p_e	p_m	$\hat{\kappa}$	SE($\hat{\kappa}$)
Proposed	$\hat{\kappa}_W$	34	0.80	0.69	0.84	0.72	0.049
Consensus (majority)	$\hat{\kappa}_{C,W1}$	32	0.88	0.71	1	0.60	0.11
Consensus (50%)	$\hat{\kappa}_{C,W2}$	18	0.93	0.60	1	0.82	0.11
Schouten	$\hat{\kappa}_{S,W}$	34	0.80	0.69	1	0.35	0.049

In Table 5.11, $\hat{\kappa}_{C,W1}$ corresponds to the consensus method using the majority rule and $\hat{\kappa}_{C,W2}$ to the 50% rule (Equation 5.35), while $\hat{\kappa}_{S,W}$ is the agreement coefficient derived by Schouten (1982). It should be noted that there were 2 items without consensus for the majority rule and 16 for the 50% rule. When calculating the mean ($\pm$ SD) of weighted kappa coefficients for all possible pairs of raters (429 pairs) between the two groups, we obtained 0.35 ± 0.06, a value similar to Schouten's index. The intraclass correlation coefficient was 0.22 ± 0.04 in the group of experts and 0.29 ± 0.03 in the group of students, reflecting a substantial heterogeneity in both groups.

5.8 Discussion

Cohen's kappa coefficient (Cohen, 1960) is widely used to measure agreement between two raters judging items on a categorical scale. Weighted (Cohen, 1968) and intraclass (Kraemer, 1979) versions of the coefficient were also proposed. Further, the method was extended to several raters (Fleiss, 1981) and to an isolated rater and a group of raters (Vanbelle and Albert, 2009a). The problem of assessing the agreement between two groups of raters is not new. Applications are numerous (e.g., van Hoeij et al. (2004); Raine et al. (2004)) and a variety of methods has been proposed over the years to deal with this problem. Several recent articles from the applied field (e.g. Kraemer et al. (2004)), however, while emphasing the importance and relevance of the problem, claim that existing solutions are not quite appropriate and that there is a need for novel and improved methods.

The usual way to solve the problem of agreement between two groups of raters is to define a consensus in each group and to quantify the agreement between them. The problem is then reduced to the case of computing Cohen's kappa agreement coefficient between two raters on a categorical scale. The rule of consensus may be defined as choosing for each item the modal (or majority) category or the category whose frequency exceeds a given percentage (e.g. 50% or 80%) in each group of raters. The consensus method, however, has serious limitations that weaken its use in practice. Indeed, a consensus is not always possible for all items (as illustrated by the SCT data) resulting in a loss of items and hence of statistical precision. The variability of the responses within each group of raters is completely ignored and the strength of the consensus is not really reflected. Further, the conclusions can be highly dependent on which definition is used for the consensus Kraemer et al. (2004). Moreover, since items without consensus (i.e., with high variability among the raters) are generally discarded from the analysis, the results obtained are prone to bias and over-optimistic estimation (see SCT example). Another natural method for assessing the concordance between two sets of raters consists in calculating the mean kappa coefficient between all possible pairs of raters composed by one rater of each group. As seen in the SCT example, this approach gives a value similar to the index developed by Schouten (1982) in the context of hierarchical clustering of raters within a single population of raters.

The agreement between two groups of raters raises the basic question of what it meant by "perfect agreement" between two groups. While this issue is meaningless in the case of two raters (they agree or they don't agree), it becomes critical at the group level agreement. The consensus method is one way to circumvent the difficulty and the mean of all pairwise kappa coefficients in another way. Schouten (1982) eluded the problem by defining perfect agreement between two groups as the situation where all raters of each group perfectly agree on all items, quite an extreme assumption. The novelty of the method derived by Vanbelle and Albert (2009b) is that it rests on a less stringent definition of perfect agreement in a population-based context. Specifically, two populations of raters are defined to be in perfect agreement (kappa coefficient equal to 1) if they have the same probability of classifying each item on the K-categorical scale. With this definition in mind, it does not really matter which raters agree or don't agree for a given item within each population, as long as the proportions in the two populations are equal. Each population is viewed as a global entity with its own heterogeneity and there is no direct interest in the agreement of individual raters within or between populations. Actually, it is quite possible that the two populations perfectly agree while a substantial part of raters disagree with each other in their own population and with some raters in the other population. As a consequence of the definition of perfect agreement, the maximum attainable proportion of agreement between the two

populations (at least in the dichotomous case) can be expressed as an analytical function of two factors, the intraclass correlation coefficient within each population and the overall marginal probabilities of classifying the items. By setting the intraclass correlation coefficient equal to 1, it turns out that the approach of Vanbelle and Albert (2009b) rejoins Schouten's assumption of perfect agreement, which can therefore be regarded as a special (extreme) case of their general definition. As illustrated on the SCT data, the difference between Schouten's and Vanbelle and Albert (2009b) approach can be marked ($\hat{\kappa} = 0.72$ and 0.35, respectively). This is due to the fact that both groups of raters show a high variability in their responses (the ICC was 0.22 ± 0.04 in the group of experts and 0.29 ± 0.03 in the group of students, respectively). The method of Vanbelle and Albert (2009b) allows for prefect agreement in presence of group heterogeneity while Schouten's approach does not. Schouten's index, however, can be derived directly from the $K \times K$ contingency table of joint probabilities estimates, whereas this is not possible with the proposed approach because the definition of perfect agreement requires the raw original data to be available to compute the maximum attainable value. As for the sampling variability aspects, Vanbelle and Albert (2009b) suggested to use the Jackknife method rather than by asymptotic formulas.

The agreement index proposed by Vanbelle and Albert (2009b) is also superior to the consensus approach (a method that we tried to formalize more theoretically) in the sense that it takes into account the variability among raters in each population and it incorporates always all items to be allocated. An intraclass and weighted versions were also proposed. If there is only one rater in each group, all coefficients envisaged here reduce to Cohen's kappa coefficient. Recently, Vanbelle and Albert (2009a) envisaged the agreement between a single rater and a group of raters, a situation which may be regarded as a special case of the present one but which raises specific problems in practice.

In conclusion, the index proposed by Vanbelle and Albert (2009b) measures the overall agreement between two independent groups of raters, taking into account the within group heterogeneity. The method is a natural extension of Cohen's kappa coefficient and demonstrates similar properties.

5.9 Proofs

Equivalence 5. *We have*

$$\kappa = \frac{\sum_{j=1}^{K}(\Pi_{[j]T} - \Pi_{[j]E})}{\sum_{j=1}^{K}(\Pi_{[j]M} - \Pi_{[j]E})} = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E}$$

where the quantities $\Pi_{[j]T}$, $\Pi_{[j]E}$ *and* $\Pi_{[j]M}$ *correspond to the quantities described in the dichotomous case when the nominal scale is dichotomized by grouping all categories other than category* j *together and* Π_T, Π_E *and* Π_M *are defined by*

$$\Pi_T = \sum_{j=1}^{K} E(P_{ij,1}P_{ij,2}); \quad \Pi_E = \sum_{j=1}^{K} \pi_{j,1}\pi_{j,2}; \quad \Pi_M = \sum_{j=1}^{K} E(P_{ij}^2).$$

Proof. Indeed, when grouping all categories other than $[j]$ together, a 2×2 table cross-classifying populations of raters $\mathcal{R}_1$ and $\mathcal{R}_2$ with respect to category j of the nominal scale can be constructed ($j = 1, \cdots, K$) (Table 5.12).
Thus,

$$\begin{aligned}
\sum_{j=1}^{K} \Pi_{[j]T} &= \sum_{j=1}^{K} E[P_{ij,1}P_{ij,2} + (1 - P_{ij,1})(1 - P_{ij,2})]) \\
&= E(2\sum_{j=1}^{K} P_{ij,1}P_{ij,2} + \sum_{j=1}^{K} 1 - \sum_{j=1}^{K} P_{ij,1} - \sum_{j=1}^{K} P_{ij,2}) \\
&= 2E(\sum_{j=1}^{K} P_{ij,1}P_{ij,2}) + K - 2 \\
&= 2\Pi_T + K - 2.
\end{aligned}$$

Table 5.12. 2×2 table cross-classifying the two populations of raters with respect to a nominal scale, obtained when grouping all categories other than category $[j]$ together

		$\mathcal{R}_2$		
		[j]	Other	
$\mathcal{R}_1$	[j]	$E[P_{ij,1}P_{ij,2}]$	$E[P_{ij,1}(1-P_{ij,2})]$	$\pi_{j,1}$
	Other	$E[(1-P_{ij,1})P_{ij,2}]$	$E[(1-P_{ij,1})(1-P_{ij,2})]$	$1-\pi_{j,1}$
		$\pi_{j,2}$	$1-\pi_{j,2}$	1

Likewise, it is easily seen that

$$\sum_{j=1}^{K} \Pi_{[j]E} = \Pi_E + K - 2 \text{ and } \sum_{j=1}^{K} \Pi_{[j]M} = \Pi_M + K - 2.$$

It follows immediately that

$$\kappa = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E}.$$

■

CHAPTER 6

Tests on agreement indexes

6.1 Introduction

Agreement indexes between two raters, several raters, an isolated rater and a group of raters and two groups of raters were introduced in previous chapters. The large sample variance of these agreement indexes was also derived and allows the determination of confidence intervals and testing if agreement is greater than obtained by chance. This chapter investigates in more detail statistical tests for a single kappa coefficient and for comparing several kappa coefficients. For the latter, we shall distinguish agreement indexes obtained on independent samples and those derived from the same sample of items. Fleiss (1981) developed a method based on the chi-square decomposition theory for comparing two or more independent agreement indexes. No method for comparing two dependent agreement indexes was available before the work of McKenzie et al. (1996), who described a resampling method based on the bootstrap. This method was generalized to several agreement indexes by Vanbelle and Albert (2008). A third type of comparison may also arise, as discussed by Williams (1976) and Schouten (1982), when testing the effect of a single rater on the agreement within a group of raters. Schouten (1982) compared the agreement obtained between an isolated rater (I) and a group of raters (G) to the agreement within the group of raters formed by the group of raters and the isolated rater (I+G) to detect raters who significantly lower the agreement. The methods exposed in this chapter will be illustrated on the comparison of agreement indexes between two raters, between an isolated rater and a group of raters and between two groups of raters.

6.2 Test on a single kappa coefficient

6.2.1 Asymptotic method

In order to test the following hypothesis for some fixed κ_0,

$$H_0 : \ \kappa = \kappa_0 \text{ vs } H_1 : \ \kappa \neq \kappa_0,$$

consider the statistic

$$Z = \frac{\hat{\kappa} - \kappa_0}{SE(\hat{\kappa})} \tag{6.1}$$

following asymptotically a Normal distribution $Z \sim N(0,1)$. H_0 is rejected if the observed Z statistic (Z_{obs}) is such that

$$\mid Z_{obs} \mid \geq Q_Z(1 - \frac{\alpha}{2}) \tag{6.2}$$

where $Q_Z(1 - \frac{\alpha}{2})$ is the $(1 - \frac{\alpha}{2})$-quantile of the Normal distribution. H_0 is not rejected otherwise.

The $(1-\alpha)100\%$ confidence interval for a kappa statistic is thus defined by

$$\hat{\kappa} - Q_Z(1 - \frac{\alpha}{2})SE(\hat{\kappa}) < \kappa < \hat{\kappa} + Q_Z(1 - \frac{\alpha}{2})SE(\hat{\kappa}).$$

6.2.2 Bootstrap method

Klar et al. (2002) proposed the use of the bootstrap method to form a $(1 - \alpha)$ percentile confidence interval for Cohen's kappa coefficient when the scale is binary. This should be interesting for small sample size ($N < 200$) because it has be shown that the kappa statistic is not symmetrically distributed in that case (Bloch and Kraemer, 1989). This is partially due to the fact that the statistic is bounded by the value 1.

Suppose that two raters classify N independent items on a binary scale and let $Y_{i,r}$ denote the binary random variable associated with the classification of the raters (see Section 2.3.3). Let $P_{jk} = P(Y_{i,1} = j, Y_{i,2} = k)$ denote the joint probabilities of $Y_{i,1}$ and $Y_{i,2}$. The joint probability function of $Y_{i,1}$ and $Y_{i,2}$ can be written as

$$P(Y_{i,1} = y_{i,1}, Y_{i,2} = y_{i,2}) = P_{11}^{y_{i,1}y_{i,2}} P_{12}^{y_{i,1}(1-y_{i,2})} P_{21}^{(1-y_{i,1})y_{i,2}} P_{22}^{(1-y_{i,1})(1-y_{i,2})}. \tag{6.3}$$

Considering all N independent items, the joint distribution of the data is multinomial

$$f(n_{11}, n_{12}, n_{21}, n_{22}|\boldsymbol{P}) = \frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!} P_{11}^{n_{11}} P_{12}^{n_{12}} P_{21}^{n_{21}} P_{22}^{n_{22}} \tag{6.4}$$

where $\boldsymbol{P} = (P_{11}, P_{12}, P_{21}, P_{22})'$, $n_{11} = \sum_{i=1}^{N} y_{i,1}y_{i,2}$, $n_{12} = \sum_{i=1}^{N} y_{i,1}(1 - y_{i,2})$, $n_{21} = \sum_{i=1}^{N}(1-y_{i,1})y_{i,2}$, $n_{22} = \sum_{i=1}^{N}(1-y_{i,1})(1-y_{i,2})$ and $n_{11}+n_{12}+n_{21}+n_{22} = N$

(see Table 2.2). The maximum likelihood estimate (MLE) of Cohen's kappa coefficient for binary scales is given by Equation 2.22.

The bootstrap method first consists in creating a finite population, obtained by giving each observation in the data set a probability of $1/N$. From this finite population, there are N^N possible samples of size N obtained with replacement. For each of these samples, the corresponding MLE of Cohen's kappa coefficient can be calculated, giving N^N estimates. The empirical distribution of these N^N estimates of Cohen's kappa coefficient is the *exact bootstrap* distribution.

Consider a random draw, say $(Y_{i,1}^*, Y_{i,2}^*)$ from the finite population. Since $(Y_{i,1}^*, Y_{i,2}^*)$ can take only four possible values, using discrete probability theory, we have

$$P(Y_{i,1}^* = y_1, Y_{i,2}^* = y_2) = p_{11}^{y_1 y_2} p_{12}^{y_1(1-y_2)} p_{21}^{(1-y_1)y_2} p_{22}^{(1-y_1)(1-y_2)} \tag{6.5}$$

where $p_{jk} = n_{jk}/N$, $j, k = 1, 2$. Next, consider all N^N possible samples of size N. Since there are only four possible values of (y_1, y_2), many of these samples will be identical. In particular, the probability of obtaining a sample of size N in which m_{11} of the $(Y_{i,1}^*, Y_{i,2}^*)$s equal $(1, 1)$, m_{12} of the $(Y_{i,1}^*, Y_{i,2}^*)$s equal $(1, 0)$, m_{21} of the $(Y_{i,1}^*, Y_{i,2}^*)$s equal $(0, 1)$ and m_{22} of the $(Y_{i,1}^*, Y_{i,2}^*)$s equal $(0, 0)$, is

$$f(m_{11}, m_{12}, m_{21}, m_{22}|\boldsymbol{p}) = \frac{N!}{m_{11}!m_{12}!m_{21}!m_{22}!} p_{11}^{m_{11}} p_{12}^{m_{12}} p_{21}^{m_{21}} p_{22}^{m_{22}} \tag{6.6}$$

where $m_{22} + m_{21} + m_{12} + m_{11} = N$ and $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22})'$. Thus, instead of needing to calculate explicitly all N^N samples, the sample (i.e. all possible values) of $(m_{11}, m_{12}, m_{21}, m_{22})$ is sufficient to determine the exact bootstrap distribution of Cohen's kappa coefficient. To calculate the number of points in the sample space Klar et al. (2002) supposed that m_{11} vary from 0 to N; then m_{12} can vary from 0 to $N - m_{11}$ and m_{21} can vary from 0 to $N - m_{12} - m_{11}$. The total number of points in the sample space is thus

$$\sum_{m_{11}=0}^{N} \sum_{m_{12}=0}^{N-m_{11}} \sum_{m_{21}=0}^{N-m_{12}-m_{11}} 1 = \frac{(N+3)(N+2)(N+1)}{6}. \tag{6.7}$$

For each sample point $(m_{11}, m_{12}, m_{21}, m_{22})$, the value of Cohen's kappa coefficient and its associated probability, given by Equation 6.6, are determined.

To obtain a two-sided $(1 - \alpha)100\%$ confidence interval, Klar et al. (2002) ordered the $(N + 3)(N + 2)(N + 1)/6$ values of Cohen's kappa coefficient to calculate the bootstrap distribution function and determine the percentiles $\alpha/2$ and $(1 - \alpha/2)$. Since the bootstrap distribution is discrete, it is very unlikely that the standard percentiles will occur at points of the discrete distribution. As such, Klar et al.

(2002) suggested linear interpolation to calculate the percentiles. For the Qth percentile, if Q_U is the closest percentile greater than Q (with the corresponding Cohen's kappa coefficient equal to $\hat{\kappa}_U$) and Q_L is the closest percentile less than Q (with the corresponding Cohen's kappa coefficient equals to $\hat{\kappa}_L$), then the Qth percentile obtained using linear interpolation is

$$\hat{\kappa}_Q = \frac{\hat{\kappa}_L(Q_U - Q) + \hat{\kappa}_U(Q - Q_L)}{Q_U - Q_L}. \tag{6.8}$$

Instead of using linear interpolation, Klar et al. (2002) proposed to choose the closest observed percentile that is less than $(\alpha/2)$ and the closest observed percentile that is greater than $(1 - \alpha/2)$. This is called the *conservative method*. After using linear interpolation or the conservative method, confidence interval can further be refined by using the *bias-corrected* method (Efron and Tibshirani, 1993).

Finally, the exact bootstrap estimate of Cohen's kappa can be calculated as

$$\overline{\kappa} = \sum_{m_{11}=0}^{N} \sum_{m_{12}=0}^{N-m_{11}} \sum_{m_{21}=0}^{N-m_{12}-m_{11}} f(m_{11}, m_{12}, m_{21}, m_{22}|\boldsymbol{p})\hat{\kappa}(m_{11}, m_{12}, m_{21}, m_{22}).$$

Then, $\overline{\kappa}$ can be used to calculate the exact bootstrap estimate of the variance,

$$var(\hat{\kappa}) = \sum_{m_{11}=0}^{N} \sum_{m_{12}=0}^{N-m_{11}} \sum_{m_{21}=0}^{N-m_{12}-m_{11}} f(m_{11}, m_{12}, m_{21}, m_{22}|\boldsymbol{p})(\hat{\kappa}(m_{11}, m_{12}, m_{21}, m_{22}) - \overline{\kappa})^2$$

and the exact bootstrap estimate of bias,

$$BIAS = \hat{\kappa}_{ML} - \overline{\kappa}.$$

6.3 Tests on independent kappas

Independent agreement indexes refer to agreement indexes obtained on different populations of items. This is the case when determining the agreement in the G modalities of a categorical covariate. The raters may be the same or different in the G modalities. For example, agreement between the two same raters may be quantified for men and women ($G = 2$).

6.3.1 Two kappa coefficients

Suppose that we have to compare the agreement between raters obtained for two independent samples of items. Let $\hat{\kappa}_1$ and $\hat{\kappa}_2$ be the two kappa coefficients, respectively. To test the hypotheses $H_0: \kappa_1 = \kappa_2$ vs $H_1: \kappa_1 \neq \kappa_2$, the statistic

$$|Z| = \frac{\hat{\kappa_1} - \hat{\kappa}_2}{\sqrt{var(\hat{\kappa}_1) + var(\hat{\kappa}_2)}} = \frac{\hat{\kappa}_1 - \hat{\kappa}_2}{SE(\hat{\kappa}_1) + SE(\hat{\kappa}_2)} \tag{6.9}$$

follows asymptotically a Normal distribution $Z \sim N(0,1)$. H_0 is rejected if the observed Z statistic (Z_{obs}) is such that

$$\mid Z_{obs} \mid \geq Q_Z(1-\frac{\alpha}{2}) \tag{6.10}$$

where $Q_Z(1-\frac{\alpha}{2})$ is the $(1-\frac{\alpha}{2})$-quantile of the Normal distribution. H_0 is not rejected otherwise.

6.3.2 Several kappa coefficients

Fleiss (1981) developed a method directly inspired by the classical one-way analysis of variance and the chi-square decomposition theory for comparing several association measures. This methodology is applied to the kappa coefficients in this section. Consider G independent estimates of a kappa coefficient $(\hat{\kappa}_1, \cdots, \hat{\kappa}_G)$. The coefficient $\hat{\kappa}_g$ denotes the kappa coefficient relative to modality g of the categorical covariate $(g = 1, \cdots, G)$. Let $SE(\hat{\kappa}_g)$ be the standard error of the kappa coefficient $\hat{\kappa}_g$ and $w_g = 1/[SE(\hat{\kappa}_g)]^2$. Under the hypothesis of agreement only due to chance in the modality g of the covariate, the statistic

$$\chi_g = \frac{\hat{\kappa}_g}{SE(\hat{\kappa}_g)} = \hat{\kappa}_g\sqrt{w_g} \tag{6.11}$$

follows approximately a Normal distribution (central-limit theorem) and the statistic

$$\chi_g^2 = w_g\hat{\kappa}_g^2 \tag{6.12}$$

follows approximately a chi-square distribution with one degree of freedom if the sample sizes n_g $(g = 1, \cdots, G)$ are sufficiently 'large'. Fleiss (1981) considered the statistic

$$\chi_{tot}^2 = \sum_{g=1}^{G} \chi_g^2 \tag{6.13}$$

to compare the G kappa coefficients. Under the hypothesis of no agreement in each of the G modalities, χ_{tot}^2 follows a chi-square distribution with G degrees of freedom.

Fleiss (1981) divided χ_{tot}^2 in two terms $\chi_{tot}^2 = \chi_{hom}^2 + \chi_{ass}^2$ where χ_{hom}^2 represents the homogeneity degree between the G kappa coefficients and χ_{ass}^2 represents a mean degree of agreement. The term χ_{ass}^2 is computed as followed,

$$\hat{\kappa}_{ass} = \frac{\sum_{g=1}^{G} w_g\hat{\kappa}_g}{\sum_{g=1}^{G} w_g}. \tag{6.14}$$

Under the hypothesis of a global kappa coefficient equal to 0, $\hat{\kappa}_{ass}$ has a value of 0 and

$$SE(\hat{\kappa}_{ass}) = \frac{1}{\sqrt{\sum_{g=1}^{G} w_g}}.$$

$$Z_{ass} = \frac{\hat{\kappa}_{ass}}{SE(\hat{\kappa}_{ass})} = \frac{\sum_{g=1}^{G} w_g \hat{\kappa}_g}{\sqrt{\sum_{g=1}^{G} w_g}}$$

is thus normally distributed. Fleiss (1981) considered that the statistic defined by

$$\chi^2_{ass} = Z^2_{ass} = \hat{\kappa}^2_{ass} \sum_{g=1}^{G} w_g = \frac{\left(\sum_{g=1}^{G} w_g \hat{\kappa}_g\right)^2}{\sum_{g=1}^{G} w_g}$$

follows approximately a chi-square distribution with one degree of freedom. The term χ^2_{hom} is obtained by subtraction.

$$\chi^2_{hom} = \chi^2_{tot} - \chi^2_{ass} = \sum_{g=1}^{G} w_g \hat{\kappa}^2_g - \hat{\kappa}^2_{ass} \sum_{g=1}^{G} w_g = \sum_{g=1}^{G} \frac{(\hat{\kappa}_g - \hat{\kappa}_{ass})^2}{[SE(\hat{\kappa}_g)]^2}. \tag{6.15}$$

In order to test the hypothesis $H_0 : \kappa_1 = \cdots = \kappa_G$ vs $H_1 : \exists\, i \neq j : \kappa_i \neq \kappa_j$ $(i, j \in \{1, \cdots, G\})$, we have to compare χ^2_{hom} to the chi-square distribution with $G-1$ degrees of freedom, the null hypothesis being rejected at the α confidence level if χ^2_{hom} is greater than $Q_{\chi^2}(1-\alpha; G-1)$, the $(1-\alpha)$-quantile of the chi-square distribution on $G-1$ degrees of freedom. The expression $[SE(\hat{\kappa}_g)]^2$ was originally derived from the Delta method (see Chapter 2).

6.4 Test on dependent kappas

Dependent agreement indexes are obtained by determining an agreement index several times on the same population of items. The raters may be the same or different. For example, the agreement between two raters on a sample of items using a first method may be compared with the agreement index obtained on the same sample of items with a second method.

6.4.1 Selection of homogeneous subgroups of raters

Schouten (1982) developed a method to test whether removing one rater from a group of raters significantly increased the agreement in the group of raters. Consider a group of $R+1$ raters. Let $\hat{\kappa}_r(w)$ designate Schouten's measure of agreement between rater r and the R remaining raters in the group and $\hat{\kappa}(w)$ designate Schouten's group measure of agreement between the $R+1$ raters, Schouten (1982)

showed that $\hat{\kappa}(w)$ significantly increases by removing the rater r from the group if

$$\chi^2_{(1)} = \frac{(\hat{\kappa}(w) - \hat{\kappa}_r(w))^2}{var(\hat{\kappa}(w)) + var(\hat{\kappa}_r(w)) - 2cov(\hat{\kappa}(w), \hat{\kappa}_r(w))} \tag{6.16}$$

is greater than the $(1-\alpha)$-quantile of the chi-square statistic on one degree of freedom. This permits to study whether a rater is an "outlier" by significantly decreasing the agreement existing in the group of raters. It is suggested to determine the large sample variance and the large sample covariance using the Jackknife technique.

6.4.2 Two kappa coefficients

Suppose that raters classify N items on a categorical scale at two different occasions or in two different experimental settings. Let $\hat{\kappa}_1$ and $\hat{\kappa}_2$ be the agreement indexes obtained. Since the two agreements are assessed on the same items, $\hat{\kappa}_1$ and $\hat{\kappa}_2$ are correlated. Are they statistically different? Let $H_0 : \kappa_1 = \kappa_2$, the null hypothesis to be tested. The bootstrap method consists in drawing q samples (1000 is generally sufficient following McKenzie et al. (1996)) of size N with replacement. For each generated sample, the agreement coefficient is estimated in the two settings and their difference $\widehat{\kappa}_d = \widehat{\kappa}_2 - \widehat{\kappa}_1$ calculated. McKenzie et al. (1996) suggested to determine the bootstrap two-sided $(1-\alpha)100\%$ confidence interval for the $\hat{\kappa}_d$ differences, whence rejecting the null hypothesis if the confidence interval does not include 0. This approach is equivalent to using a Student's t-test (Vanbelle and Albert, 2008) and to reject H_0 at the α significance level if

$$|t_{obs}| = \left| \frac{\overline{\kappa}_d}{SE(\hat{\kappa}_d)} \right| \geq Q_t(1 - \frac{\alpha}{2}; q-1) \tag{6.17}$$

where $\overline{\kappa}_d$ and $SE(\hat{\kappa}_d)$ are respectively the mean and standard deviation of the q bootstrapped kappa differences and $Q_t(1 - \frac{\alpha}{2}; q-1)$ is the upper $(\alpha/2)$-quantile of the Student's t distribution on $q-1$ degrees of freedom. Otherwise, H_0 is not rejected.

McKenzie et al. (1996) proposed alternatively to use a Monte Carlo permutation test, consisting in shuffling the sample 999 times. The number of times that the difference between the original values of the agreement indexes is equaled or exceeded by the difference between the randomly permuted values is then obtained $(\hat{\kappa}_c)$. This value, incremented by one is divided by 1000.

$$\hat{\kappa}_p = \frac{\hat{\kappa}_c + 1}{1000}.$$

If the resulting value $\hat{\kappa}_p$ is less than or equal to the α significance level, then the null hypothesis is able to be rejected.

6.4.3 Several kappa coefficients

Vanbelle and Albert (2008) generalized the method of McKenzie et al. (1996) to the comparison of more than two agreement indexes. Suppose we want to compare $G \geq 2$ correlated agreement indexes $(\kappa_1, \cdots, \kappa_G)$, i.e., to test the null hypothesis $H_0 : \kappa_1 = \cdots = \kappa_G$ against the alternative hypothesis $H_1 : \exists k \neq l \in \{1, \cdots, G\} : \kappa_k \neq \kappa_l$. As before, the bootstrap method will consist in drawing q samples of size N with replacement from the original data. Then, for each bootstrapped sample $(j = 1, \cdots, q)$, let $\widehat{\boldsymbol{\kappa}}_j = (\widehat{\kappa}_{1(j)}, \cdots, \widehat{\kappa}_{G(j)})'$ be the vector of the G agreement coefficients obtained. The null and alternative hypotheses can be rewritten in matrix form as follows: $H_0 : \boldsymbol{C\kappa} = \mathbf{0}$ versus $H_1 : \boldsymbol{C\kappa} \neq \mathbf{0}$, where $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_G)'$ and $\boldsymbol{C}$ the $(G-1) \times G$ patterned matrix

$$\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

Then, the test statistic is

$$T^2 = (\boldsymbol{C}\overline{\boldsymbol{\kappa}})'(\boldsymbol{CSC}')^{-1}\boldsymbol{C}\overline{\boldsymbol{\kappa}}, \tag{6.18}$$

distributed as Hotelling's T^2, where $\overline{\boldsymbol{\kappa}}$ and $\boldsymbol{S}$ are respectively the sample mean vector and covariance matrix of the q bootstrapped vectors $\hat{\boldsymbol{\kappa}}$. The null hypothesis will be rejected at the α confidence level if

$$T^2 \geq \frac{(q-1)(G-1)}{(q-G+1)} Q_F(1-\alpha; G-1, q-G+1) \tag{6.19}$$

where $Q_F(1-\alpha; G-1, q-G+1)$ is the upper α-percentile of the F distribution on $G-1$ and $q-G+1$ degrees of freedom. Otherwise, H_0 will not be rejected. Note that, since "$q-G+1$" will be large in general, the left-hand side of Equation 6.19 can be approximated by $Q_{\chi^2}(1-\alpha; G-1)$, the $(1-\alpha)$ percentile of the chi-square distribution on $G-1$ degrees of freedom. If $\boldsymbol{c}_g$ denotes the g-th row of matrix $\boldsymbol{C}$, simultaneous confidence intervals for individual contrasts $\boldsymbol{c}_g'\boldsymbol{\kappa}$ $(g = 1, \cdots, G-1)$ given by

$$\boldsymbol{c}_g'\overline{\boldsymbol{\kappa}} \pm \sqrt{\frac{(q-1)(G-1)}{(q-G+1)} Q_F(1-\alpha; G-1, q-G+1)} \sqrt{\boldsymbol{c}_g'\boldsymbol{S}\boldsymbol{c}_g} \tag{6.20}$$

can be used for multiple comparison purposes. When $G = 3$, Vanbelle and Albert (2008) proposed to represent graphically the data with a 95% confidence ellipse for the differences in agreement.

6.5 Examples

6.5.1 Blood clots detection

The presence of blood clots was assessed on 50 patients (23 women and 27 men) with a reference method (Standard) and two new methods (Method 1 and Method 2) by two medical raters (see Chapter 2, Section 2.6.2). For each new method, Cohen's kappa coefficients were determined for men and women and compared with the method developed by Fleiss (1981) (see Table 6.1). The Cohen's kappa coefficient was 0.27 ± 0.19 ($p = 0.16$) for men and 0.47 ± 0.16 ($p = 0.0034$) for women with Method 1 and 0.57 ± 0.17 ($p = 0.0008$) for men and 0.83 ± 0.12 ($p < 0.0001$) for women with Method 2. The agreement obtained for men with Method 1 was not better than chance while all other agreement indexes were greater than chance. For both methods, Cohen's kappa coefficients for men and women were homogeneous.

Table 6.1. Results of the chi-square test comparing Cohen's kappa coefficients obtained for men and women when detecting blood clots with a new method (Method 1 and Method 2) and a reference method

	Men (N=27)	Women (N=23)	$\hat{\kappa}_{ass}$	χ^2	p-value
Method 1	0.27 ± 0.19	0.47 ± 0.16	0.39	0.62	0.43
Method 2	0.57 ± 0.17	0.83 ± 0.12	0.74	1.52	0.22

Then, Cohen's kappa coefficient obtained using Method 1 ($\hat{\kappa} = 0.41 \pm 0.12$) was compared to the kappa coefficient obtained using Method 2 ($\hat{\kappa} = 0.71 \pm 0.10$) using the bootstrap method with 2000 iterations and the permutation test. Figure 6.1 shows the different kappa values drawn with the bootstrap and the permutation methods.

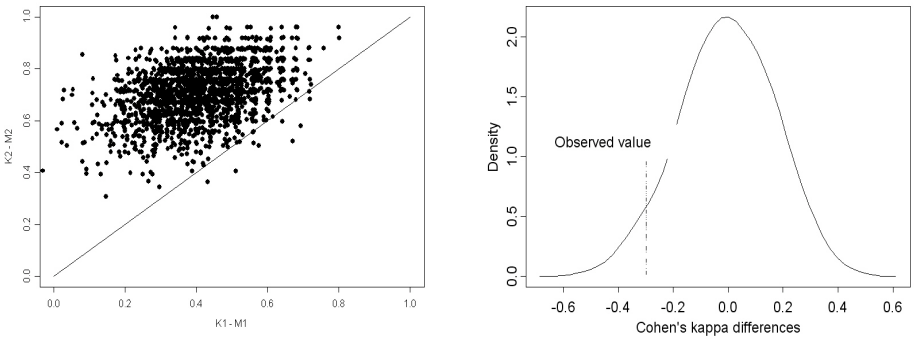

Figure 6.1. Kappa values drawn by the bootstrap (left) and densities of Cohen's kappa differences obtained with the permutation test (right) for the comparison of kappa coefficients for Method 1 and Method 2 when detecting blood clots

The estimated correlation between the two Cohen's kappa coefficients was 0.41 ($p < 0.0001$). The bootstrap method (Vanbelle and Albert, 2008) yielded a significant result ($p = 0.018$) and the permutation test (McKenzie et al., 1996) a 95% confidence interval of $[-0.54, -0.058]$. Method 2 should thus be preferred to Method 1 because there is a better agreement with the reference method.

6.5.2 Cervical ectopy size

Cervical ectopy size of 85 women was determined on a 4-category scale by two medical raters by direct visual assessment and with the computerized planimetry method. To test if agreement between the two raters is the same with the planimetry and the visual method, the methods developed by McKenzie et al. (1996) and Vanbelle and Albert (2008) were used. Remember that the weighted kappa coefficient (with quadratic weights) was 0.67 ± 0.062 for direct visual assessment and 0.82 ± 0.053 for the planimetry method. Both agreement indexes were better than chance ($p < 0.0001$). The bootstrap method with 2000 iterations led to [-0.29,-0.0023] as 95 % confidence interval for the weighted kappa differences and to a p-value of $p = 0.030$ for the Student t-test. The estimated correlation between the two kappas was 0.19 (p<0.0001). Using the permutation test, the p-value was $p = 0.031$ for weighted kappa coefficient differences. Therefore, the planimetry method should be preferred to visual assessment since the agreement between the two raters was better. The results of the bootstrap steps and the density of the kappa differences obtained by the permutation test are given in Figure 6.2.

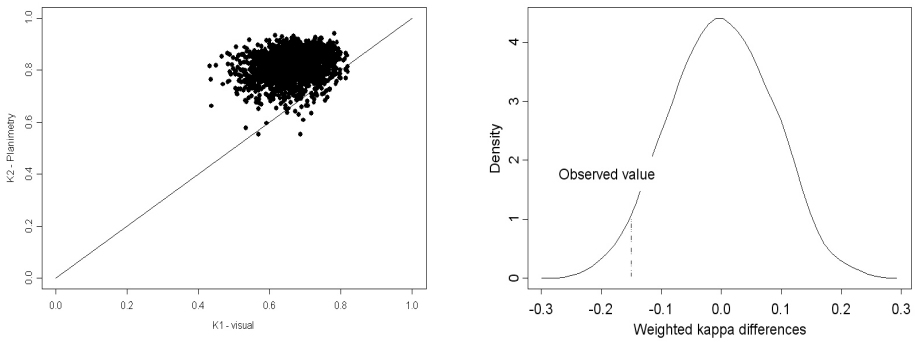

Figure 6.2. Kappa values drawn by the bootstrap (left) and densities of the kappa differences obtained from the permutation test (right) for the quadratic weighted kappa coefficient obtained between two medical raters with direct visual assessment and the planimetry method in the assessment of the cervical ectopy size of 85 women

6.5.3 Deep venous thrombosis

A study was conducted on 107 patients in the medical imaging department of the university hospital to compare deep venous thrombosis (DVT) detection using a multidetector-row computed tomography (MDCT) and ultrasound (US) (Vanbelle and Albert, 2008). The study also looked at the benefit of using spiral (more images and possibility of multiplanar reconstructions) with respect to sequential technique (less slices, less irradiation). Images were acquired in the spiral model (ankle to inferior vena cava) and reconstructed in 5 mm thickness slices every 5 mm, 20 mm and 50 mm. Two radiologists (one junior and one senior) assessed for each patient and each experimental setting (5/5, 5/20 and 5/50 slices) the presence of DVT. The aim of the study was to compare agreement of the different MDCT slices with the US method. Only data of the senior radiologist will be presented here (see Table 6.2).

Table 6.2. Cross-classification of DVT detection (0=absence, 1=presence) using different MDCT slices (5/5, 5/20 and 5/50 mm) and US in 107 patients by a senior radiologist

	MDCT slices						
	5/5 mm		5/20 mm		5/50 mm		
US	0	1	0	1	0	1	Total
0	96	1	95	2	96	1	97
1	0	10	1	9	2	8	10
Total	96	11	96	11	98	9	107
	$\hat{\kappa}_{5/5} = 0.95$		$\hat{\kappa}_{5/20} = 0.84$		$\hat{\kappa}_{5/50} = 0.83$		

The observed Cohen's kappa coefficients ($\pm SE$) were 0.95 ± 0.053, 0.84 ± 0.089 and 0.83 ± 0.098 for 5/5, 5/20 and 5/50 mm slices, respectively. The bootstrap approach with 2000 iterations led to a Hotelling's T^2 value of 1.46 ($p = 0.48$) indicating no evidence of a difference between the κ coefficients at the 5% significance level. The bootstrap estimates of bias were 0.003, 0.008 and 0.009 for the 5/5, 5/20 and 5/50 mm slices, respectively. According to the rule described in Efron and Tibshirani (1993), the bias can be ignored. The differences between the Cohen's kappa coefficients generated by the 2000 iterations of the bootstrap are represented in Figure 6.3 with the 95% confidence ellipse for the difference vector ($\hat{\kappa}_{5/5} - \hat{\kappa}_{5/20}$, $\hat{\kappa}_{5/5} - \hat{\kappa}_{5/50}$). It is seen that the origin $(0, 0)$ is well inside the confidence region, as expected.

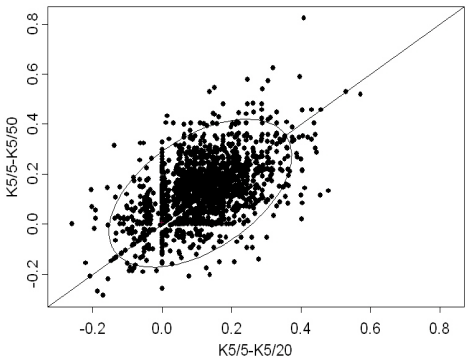

Figure 6.3. Kappa differences ($\hat{\kappa}_{5/5} - \hat{\kappa}_{5/50}$ vs $\hat{\kappa}_{5/5} - \hat{\kappa}_{5/20}$) generated by the bootstrap ($q = 2000$) with 95% CI

6.5.4 Script Concordance Test

During 2006, a SCT in endocrinology was proposed to students in medicine in their 3, 4, 5 or 6th year at the University of Liège, Belgium (Collard et al., 2009). The SCT consisted of 48 items relating possible situations encountered in endocrinology. There were 35, 20, 26 and 27 students passing the SCT in the 3, 4, 5 and 6th study year, respectively. Their responses were confronted to the responses given by a panel of 10 experts. The 48 items were divided in two categories: situations encountered by the students during their study ("inside context") and situations never seen during lessons ("outside context"). Firstly, the agreement obtained between each expert and the remaining 9 experts in the panel was compared with the agreement in the panel of experts using the method of Schouten (1982) to see if some experts significantly decrease the agreement among the panel. Using linear weights, the weighted agreement index derived by Schouten (see Chapter 4, Section 4.5) within the panel of experts was equal to 0.49 ± 0.034. Expert 6 appeared to lower significantly the agreement in the panel of expert but was left in the study (see Table 6.3).

Table 6.3. Results of Schouten's method of homogeneous subgroups selection when considering the group of 10 experts in the SCT example

Expert	1	2	3	4	5	6	7	8	9	10
$\hat{\kappa}_r(w)$	0.50	0.52	0.54	0.53	0.48	0.34	0.52	0.49	0.50	0.42
SE	0.055	0.046	0.041	0.045	0.051	0.079	0.042	0.052	0.044	0.046
χ^2	0.14	1.78	4.28	2.36	0.024	5.49	1.12	0.040	0.21	2.98
p-value	0.71	0.18	0.039	0.12	0.88	0.019	0.29	0.84	0.64	0.084

Then, to study if the agreement between the students and the panel of experts was related to the level of education (i.e., years of studies), the agreement between the group of students and experts was determined in each study year (Vanbelle and Albert, 2009b) and these agreement indexes were compared using the bootstrap method (Vanbelle and Albert, 2008). The level of agreement between each group of students and the panel of experts was determined using a linearly weighted agreement index and are given in Table 6.4 and displayed in Figure 6.4.

Table 6.4. Linearly weighted agreement indexes between the 108 students and the panel of experts according to the study year for the SCT in endocrinology

Year	R_1	R_2	$\hat{\kappa}_w$	SE
3	10	35	0.65	0.038
4	10	20	0.64	0.049
5	10	26	0.75	0.032
6	10	27	0.72	0.033
ALL	10	108	0.71	0.033

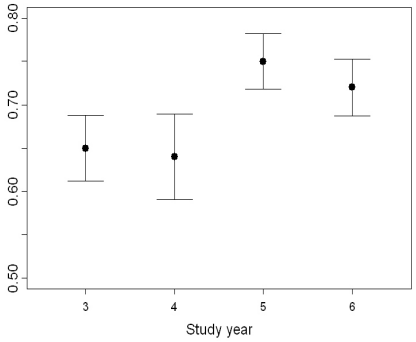

Figure 6.4. Linearly weighted agreement indexes ($\pm SE$) between the 108 students and the panel of experts according to the study year for the SCT in endocrinology

There was a significant difference according to the study year ($T^2 = 15.6$, $p = 0.0010$). Students from year 5 presented better agreement with the panel of experts than students from year 3 and 4. The estimated correlation matrix between the

agreement indexes was

$$R = \begin{pmatrix} 1 & 0.60 & 0.66 & 0.52 \\ 0.60 & 1 & 0.69 & 0.58 \\ 0.66 & 0.69 & 1 & 0.67 \\ 0.52 & 0.58 & 0.67 & 1 \end{pmatrix}$$

Finally, the agreement obtained on items "inside context" was compared to the agreement obtained on items "outside context" with the test exposed in Section 6.3.1. The hypothesis was that agreement should be better for items "inside context" than "outside context" because the situations were encountered by the students during their studies. However, there was no difference between the two types of items (see Table 6.5).

Table 6.5. Linearly weighted agreement indexes between the 108 students and the panel of experts according to the study year for the SCT in endocrinology for items "inside context" and "outside context"

			Inside context	Outside context	
Year	R_1	R_2	$\hat{\kappa}_w \pm SE$	$\hat{\kappa}_w \pm SE$	p-value
3	10	35	0.64 ± 0.055	0.68 ± 0.053	0.65
4	10	20	0.61 ± 0.076	0.68 ± 0.068	0.52
5	10	26	0.77 ± 0.041	0.76 ± 0.049	0.85
6	10	27	0.73 ± 0.052	0.73 ± 0.045	0.99
ALL	10	108	0.70 ± 0.047	0.73 ± 0.048	0.67

6.6 Discussion

This chapter was concerned with statistical hypothesis testing on kappa coefficients. We distinguished between single kappa tests and multiple kappa comparison tests. Confidence intervals were constructed for a single agreement coefficient using an asymptotic method and the bootstrap method. One advantage of the exact bootstrap procedure is that it does not force the confidence interval to be symmetric around the estimate of the kappa coefficient as it is the case with the asymptotic method. However, investigators may reach, by chance, different conclusions with the bootstrap method if the *bootstrap distribution* is not completely specified.

Then, the methods discussed in this chapter to compare several agreement indexes allow the comparison of several kappa coefficients, agreement being searched between two raters, several raters, a group of raters and an isolated rater or two

group of raters. A further distinction has to be made between the unpaired and the paired cases. The asymptotic method of Fleiss (1981) is, to our knowledge, the only method exposed in the literature to compare several independent agreement indexes, which permits the comparison of all kinds of agreement indexes. This method is based on the chi-square decomposition theory and has the disadvantage of being asymptotic but is simple to apply. No guidelines relative to the required sample size was provided with the method. To compare two or more correlated agreement coefficients, the bootstrap method of Vanbelle and Albert (2008), extending the bootstrap method of McKenzie et al. (1996) provides an estimate of the mean and the variance-covariance matrix of correlated agreement indexes and hence a way to test their homogeneity by means of the Hotelling's T^2. These methods are computer intensive and possess the drawbacks of the resampling methods but are simple to implement. It should be noted that methods restricted to particular forms of agreement indexes were not discussed in this chapter. This includes methods developed by Donner and Eliasziw (1992), Donner and Klar (1996), Donner et al. (1996) and Donner (1998) for the comparison of independent intraclass kappa coefficients between two raters. These methods are based on the common correlation model given in Section 2.4.3 and on the chi-square statistic ("goodness of fit approach"). These methods were extended to the comparison of two dependent intraclass kappa coefficients by Donner et al. (2000) and Nam (2003). None of the techniques described in this chapter allows to study the influence of continuous covariates on agreement indexes. This orientated the researchers to the modeling methods, which are developed in the next chapter.

CHAPTER 7

Regression and kappa coefficients

7.1 Introduction

Methods for comparing several agreement indexes were introduced in Chapter 6. However, with the development of generalized linear models (GLM), researchers have focused on modeling techniques to account for categorical and continuous covariates in the determination of agreement between raters. Developments were first made on the basis of hierarchical log-linear models including two components: a first component representing the effect of chance and a second the effect of rater agreement (Tanner and Young, 1985a). They first introduced methods based on the independence model in case of Cohen's kappa coefficient between two or more raters and symmetry and quasi-independence model in case of intraclass kappa coefficients. Then, for ordinal ratings, they used a linear-by-linear baseline association model since ordinal rating scales almost always exhibit a positive association between ratings (Tanner and Young, 1985b). Log-linear models were then improved by Agresti (1988, 1992) and Becker and Agresti (1992). Graham (1995) further extended the log-linear model proposed by Tanner and Young (1985a) to the analysis of categorical covariate effects on chance corrected agreement and Basu et al. (1999) used log-linear models to estimate hierarchical weighted agreement coefficients between two raters. Finally, Perkins and Becker (2002) proposed to model the bivariate marginal responses of the raters instead of modeling the joint distribution of the raters responses using log-linear models. Agresti (1992) used log-linear models but also latent class models and Rasch models to study patterns of agreements and disagreements. Latent class models were first used by

Uebersax (1988) and more recently by Schuster and Smith (2002, 2006). Latent classes emerge form the factorial combination of the true category in which the item belongs and the ease with which raters are able to classify items into the true category. Note that Schuster (2002) also used mixture model approach to index rater agreement.

Alternatively, developments were made on the basis of logistic regression analysis. First, Coughlin et al. (1992) used logistic regression analysis by defining a dependent variable equal to 1 if the raters agree and 0 otherwise and adjusted the model for independent covariates. This approach is thus not corrected for chance. Shoukri et al. (1995) and Shoukri and Mian (1996) derived the maximum likelihood estimator of the intraclass kappa coefficient when the binary classification of the raters depends on covariates relative to items and/or raters while Barlow (1996) proposed, as alternative, the use of a conditional logistic regression model to account for one or more covariates. Later, Lipsitz et al. (2001, 2003) constructed models to account for categorical and continuous covariates permitting the comparison of independent agreement indexes between two raters.

Researchers also used generalized estimating equations (GEE) to model dependent agreement indexes with respect to continuous and categorical covariates. Thomson (2001) used one set of estimating equations to estimate agreement coefficients in various situations without giving the possibility to compare the agreement coefficients obtained. On another hand, Williamson and Manatunga (1997) first used two sets of estimating equations to test for the equality of two or more dependent inter-rater agreement coefficients when ordinal ratings are made on the same sample. Then, Williamson et al. (2000) extended the methodology to the general case of agreement between two or more raters and Gonin et al. (2000) to weighted agreement indexes between two raters. Finally, Barnhart and Williamson (2002) adapted the weighted least-squares approach for comparing several dependent agreement indexes between two raters. Although various modeling techniques were developed, most of them are only applicable to one particular form of the kappa coefficient or does not link directly the agreement index to covariates. Therefore, this chapter will be limited to the empirical methods of Lipsitz et al. (2001, 2003), the weighted least-squares approach and the generalized estimating equations. These methods will be reviewed, illustrated and compared to the bootstrap method of Vanbelle and Albert (2008).

7.2 Independent agreement indexes

Lipsitz et al. (2001) developed a method for modeling Cohen's kappa coefficient as a function of covariates relative to the raters and/or the items. They proposed to

use two logistic regressions and a linear regression for binary variables to estimate Cohen's kappa coefficient. A modified version of their method was developed by Lipsitz et al. (2003).

7.2.1 Initial method

Binary scale. Let two raters assess each of N independent items on a binary scale. As before, let $Y_{i,r}$ denote the random variable such that $Y_{i,r} = 1$ if item i is rated positive by rater r and $Y_{i,r} = 0$ otherwise ($r = 1, 2$). Suppose each item has an item specific covariate vector $\boldsymbol{x}_i$ and two vectors of covariates specific to the raters $\boldsymbol{x}_{i,r},\ r = 1, 2$ as it was supposed by Shoukri and Mian (1996) and denote $\boldsymbol{z}_i' = (\boldsymbol{x}_i', \boldsymbol{x}_{i,1}', \boldsymbol{x}_{i,2}')$. Then define the indicator random variable Y_i such that $Y_i = 1$ if both raters agree on item i and $Y_i = 0$ otherwise. In terms of $Y_{i,1}$ and $Y_{i,2}$, we have

$$Y_i = Y_{i,1}Y_{i,2} + (1 - Y_{i,1})(1 - Y_{i,2}). \tag{7.1}$$

Lipsitz et al. (2001) defined Cohen's kappa coefficient between $Y_{i,1}$ and $Y_{i,2}$ as usual

$$\kappa_i = \frac{P_{oi} - P_{ei}}{1 - P_{ei}} \tag{7.2}$$

with

$$P_{oi} = P[Y_i = 1|\boldsymbol{z}_i] = P[Y_{i,1} = 1, Y_{i,2} = 1|\boldsymbol{z}_i] + P[Y_{i,1} = 0, Y_{i,2} = 0|\boldsymbol{z}_i] \tag{7.3}$$

and

$$P_{ei} = \pi_{i,1}\pi_{i,2} + (1 - \pi_{i,1})(1 - \pi_{i,2}) \tag{7.4}$$

where $\pi_{i,r} = P[Y_{i,r} = 1|\boldsymbol{x}_i, \boldsymbol{x}_{i,r}],\quad r = 1, 2.$

Adjustment for covariates associated with Cohen's kappa coefficient can be accomplished using the model

$$g(\kappa_i) = \boldsymbol{z}_i'\boldsymbol{\gamma}$$

where $g(.)$ is a link function to ensure that $-1 \leq \kappa_i \leq 1$ and $\boldsymbol{\gamma}$ is a vector of unknown parameters. Fisher's Z transformation might be used as link function. However, Lipsitz et al. (2001) found that this link function and the associated parameters $\boldsymbol{\gamma}$ are not easily interpretable and thus preferred use the identity link function without constraint on the kappa values in the estimation procedure,

$$\kappa_i = \boldsymbol{z}_i'\boldsymbol{\gamma},\quad i = 1, \cdots, N. \tag{7.5}$$

Lipsitz et al. (2001) remarked that

$$P_{oi} = P_{ei} + \kappa_i(1 - P_{ei}) = P_{ei} + \boldsymbol{z}_i'\boldsymbol{\gamma}(1 - P_{ei}) = P_{ei} + \boldsymbol{z}_i^{*'}\boldsymbol{\gamma} \tag{7.6}$$

where $\boldsymbol{z}_i^* = (1 - P_{ei})\boldsymbol{z}_i$. Therefore, if P_{ei} is known, the model of P_{oi} is a linear model with an identity link function, a known offset P_{ei}, a known covariate vector $\boldsymbol{z}_i^*$ and an unknown parameter vector $\boldsymbol{\gamma}$. Thus, to estimate $\boldsymbol{\gamma}$, if P_{ei} is known, the maximum likelihood based on the Bernoulli distribution of the random variables Y_i can be used. Unfortunately, P_{ei} is rarely known. Instead of using the maximum likelihood estimation based on the joint distribution of $(Y_{i,1}, Y_{i,2})$ to estimate jointly $(\pi_{i,1}, \pi_{i,2})$ and $\boldsymbol{\gamma}$, as made by Shoukri and Mian (1996), Lipsitz et al. (2001) replaced P_{ei} in Equation 7.6 by an estimate $\hat{P}_{ei}$ and estimated $\boldsymbol{\gamma}$ using a linear model. In particular, Lipsitz et al. (2001) estimated $\pi_{i,1}$ and $\pi_{i,2}$ using the logistic regression model

$$logit(\pi_{i,r}) = \boldsymbol{x}_{i,r}'\boldsymbol{\beta}_{1r} + \boldsymbol{x}_i'\boldsymbol{\beta}_{2r} \quad r = 1, 2 \tag{7.7}$$

and then estimated P_{ei} by

$$\hat{P}_{ei} = p_{i,1}p_{i,2} + (1 - p_{i,1})(1 - p_{i,2}) \tag{7.8}$$

where $p_{i,r} = \hat{\pi}_{i,r}$, $r = 1, 2$. Finally, they used the following linear model for P_{oi},

$$P_{oi} \approx \hat{P}_{ei} + (1 - \hat{P}_{ei})\boldsymbol{z}_i'\boldsymbol{\gamma}. \tag{7.9}$$

In practice, the estimation of the parameter vector $\boldsymbol{\gamma}$ involves

1. the use of logistic regressions of $Y_{i,r}$ versus $(\boldsymbol{x}_{i,r}, \boldsymbol{x}_i)$ to obtain $p_{i,r}$ $(r = 1, 2)$,
2. the estimation of the 'offset' $\hat{P}_{ei} = p_{i,1}p_{i,2} + (1 - p_{i,1})(1 - p_{i,2})$,
3. the use of a linear regression of the binary outcome Y_i versus $\boldsymbol{z}_i^*$ with a known offset $\hat{P}_{ei}$.

Note that $\hat{P}_{oi}$ should be in the range $[0, 1]$ but it might not be the case since the identity link function is used in the regression model of Cohen's kappa coefficient. However, Lipsitz et al. (2001) never found this to be true in the data they analyzed.

Lipsitz et al. (2001) remarked that, since the model for $\pi_{i,r}$ (see Equation 7.7) is not a function of $\boldsymbol{\gamma}$, the estimate of $\boldsymbol{\beta}_{1,r}$ and $\boldsymbol{\beta}_{2,r}$ $(r = 1, 2)$ will be the same, despite the model for κ_i. However, the estimate of κ_i depends on the model used to estimate the parameters $\pi_{i,r}$ and can be biased if the model is underfitted. Analyzing several samples, Lipsitz et al. (2001) found that it was preferable to introduce too much covariates to model $\pi_{i,r}$ than too less, even if covariates are not significant at the α significance level, a priori given. This method leads, however, to a small increase of the estimated standard error of $\hat{\boldsymbol{\gamma}}$ (Lipsitz et al., 2001).

Let $\boldsymbol{\beta}' = (\boldsymbol{\beta}_{11}', \boldsymbol{\beta}_{21}', \boldsymbol{\beta}_{12}', \boldsymbol{\beta}_{22}')$. Using Taylor series expansions similarly to Prentice (1988) and assuming that the models for $\pi_{i,1}$, $\pi_{i,2}$ and κ_i are correctly specified, it can be shown that $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')$ is consistent for $(\boldsymbol{\beta}', \boldsymbol{\gamma}')$ and that $N^{1/2}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})', (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})')$

has an asymptotic distribution which is multivariate Normal with mean vector $\mathbf{0}$ and variance-covariance matrix which can be consistently estimated by a robust variance estimator such as Jackknife estimator. One form of the Jackknife variance estimate is

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')' = \sum_{i=1}^{N}((\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i} - (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}'))'((\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i} - (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')) \tag{7.10}$$

where $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i}$ is the estimate of $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')$ obtained by deleting the two observations relative to item i and by recalculating both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

Categorical scale. Suppose now that the response is categorical with K categories and let $Y_{i,r}$ denote the random variable such that $Y_{i,r} = k$ if rater r classifies item i in category k ($r = 1, 2; i = 1, \cdots, N, k = 1, \cdots, K$). Extending the method described in the binary case, Lipsitz et al. (2001) proposed to

1. use an ordinal or polynomial logistic regression of $Y_{i,r}$ versus $(\boldsymbol{x}_i, \boldsymbol{x}_{i,r})$ to obtain $\hat{P}[Y_{i,r} = k|\boldsymbol{x}_i, \boldsymbol{x}_{i,r}]$ $(r = 1, 2)$,
2. form $\hat{P}_{ei} = \sum_{k=1}^{K} \hat{P}[Y_{i,1} = k|\boldsymbol{x}_i, \boldsymbol{x}_{i,1}]\hat{P}[Y_{i,2} = k|\boldsymbol{x}_i, \boldsymbol{x}_{i,2}]$,
3. use a linear regression of the binary outcome Y_i versus $\boldsymbol{z}_i^*$ with a known offset $\hat{P}_{ei}$ $(i = 1, \cdots, N)$ to obtain $\hat{\boldsymbol{\gamma}}$.

7.2.2 Two-stage logistic regression

Lipsitz et al. (2003) proposed to modify their method (Lipsitz et al., 2001) by using a two-stage logistic regression to estimate the agreement probability as a function of covariates. The introduction of the two-stage logistic regression was motivated by the following fact. Suppose that two binary observations are made completely independently on the same item. Moreover, suppose that the prevalence for classifying an item as positive (which is a covariate) is large in some sub-group and small in some others. Thus, by chance alone (since the reports are made independently), agreement appears to be related to the covariates although agreement is only due to chance. To overcome the problem, Lipsitz et al. (2003) proposed a two-stage logistic regression, for which the parameters of the model are equal to 0 when agreement is only due to chance using an 'offset', i.e., a known regression coefficient.

Consider the vector $\boldsymbol{z}_i$ of covariates and the indicator vector Y_i described by Equation 7.1 (see Section 7.2.1). Suppose that the probability of agreement is modeled with the following logistic regression model

$$logit(P_{oi}) = \boldsymbol{x}_i'\boldsymbol{\gamma}_1 + \boldsymbol{x}_{i,1}'\boldsymbol{\gamma}_2 + \boldsymbol{x}_{i,2}'\boldsymbol{\gamma}_3 = \boldsymbol{z}_i'\boldsymbol{\gamma}. \tag{7.11}$$

Lipsitz et al. (2003) developed a logistic regression model for which $\boldsymbol{\gamma} = \mathbf{0}$ when agreement is due to chance by introducing an offset.

$$logit(P_{oi}) = logit(P_{ei}) + \boldsymbol{x}_i'\boldsymbol{\gamma}_1 + \boldsymbol{x}_{i,1}'\boldsymbol{\gamma}_2 + \boldsymbol{x}_{i,2}'\boldsymbol{\gamma}_3 = logit(P_{ei}) + \boldsymbol{z}_i'\boldsymbol{\gamma}. \tag{7.12}$$

In practice, the procedure of Lipsitz et al. (2003) thus consists in

1. using a logistic regression of $Y_{i,r}$ versus $(\boldsymbol{x}_{i,r}, \boldsymbol{x}_i)$ to obtain $p_{i,r}$ $(r = 1, 2)$,
2. estimating the 'offset' $logit(\widehat{P}_{ei}) = logit[p_{i,1}p_{i,2} + (1 - p_{i,1})(1 - p_{i,2})]$,
3. using logistic regression of Y_i versus $(\boldsymbol{x}_i, \boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2})$ and a known offset $logit(\widehat{P}_{ei})$ to obtain $\hat{\boldsymbol{\gamma}}$.

Since the procedure does not involve a linear regression, as opposed to method described in Section 7.2.1, Lipsitz et al. (2003) solved the problem of constraints on $\hat{P}_{oi}$ $(\hat{P}_{oi} \in [0, 1])$. They also proposed, when raters are indistinguishable $(\pi_{i,1} = \pi_{i,2})$, to estimate jointly the marginal probabilities rather than separately. This consists in using one set of generalized estimating equations (GEE1) in step (1) rather than 2 separate logistic regressions.

For each item Lipsitz et al. (2003) proposed the following expression to determine how agreement differs from chance for any covariate pattern

$$\hat{\xi} = logit(\hat{P}_{oi}) - logit(\hat{P}_{ei}) = \boldsymbol{z}_i'\hat{\boldsymbol{\gamma}}. \tag{7.13}$$

In particular, the hypothesis H_0: agreement is due to chance, i.e., $logit(P_{oi}) - logit(P_{ei}) = 0$ versus H_1: $logit(P_{oi}) - logit(P_{ei}) \neq 0$ can be tested.

For any bounded monotone function $g(.)$, Lipsitz et al. (2003) proposed to use the estimate of

$$\xi_i^* = \frac{g[logit(P_{oi})] - g[logit(P_{ei})]}{\{\max g[logit(P_{oi})]\} - g[logit(P_{ei})]} \tag{7.14}$$

to test the hypothesis H_0. The summary measure ξ_i^* is equal to 0 if agreement is due to chance and 1 if agreement is prefect. In particular, for a given value of a, the choice of

$$g[a] = \frac{e^a}{1 + e^a}$$

yields an estimate of Cohen's kappa coefficient for each covariate pattern when evaluated at $(\hat{P}_{oi}, \hat{\pi}_{i,1}, \hat{\pi}_{i,2})$, that is,

$$\hat{\kappa}_i = \frac{\hat{P}_{oi} - \hat{P}_{ei}}{1 - \hat{P}_{ei}}. \tag{7.15}$$

Lipsitz et al. (2003) noted that the estimated variance of $\boldsymbol{\gamma}$ reported by standard statistical software for logistic regression will not be correct since the offset is

treated as known rather than estimated. Using Taylor series expansions similar to Prentice (1988) and assuming that the models for $\pi_{i,1}$, $\pi_{i,2}$ and P_{oi} are correctly specified, it can be shown that $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is consistent for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $N^{1/2}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})', (\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})']$ has an asymptotic distribution which is multivariate Normal with mean vector $\mathbf{0}$ and a variance-covariance matrix which can be consistently estimated by a robust variance estimator such as the sandwich estimator of White (1982) or the Jackknife estimator (Quenouille, 1956). The Jackknife estimate can be obtained as follows

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')' = \frac{N-1}{N} \sum_{i=1}^{N} ((\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i} - (\boldsymbol{\beta}', \boldsymbol{\gamma}'))'((\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i} - (\boldsymbol{\beta}', \boldsymbol{\gamma}')) \tag{7.16}$$

where $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i}$ is the estimate of $(\boldsymbol{\beta}', \boldsymbol{\gamma}')$ obtained by deleting the pair of ratings on item i.

Lipsitz et al. (2003) shown that the estimates obtained by their two-stage logistic regression are also the estimates of the generalized estimating equations

$$\boldsymbol{u}_{(\hat{\beta},\hat{\gamma})}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{N} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \left(\boldsymbol{U}_i - \boldsymbol{\eta}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \right) = \mathbf{0} \tag{7.17}$$

where $\boldsymbol{D}_i'$ is the block-diagonal matrix $\boldsymbol{D}_i' = \partial \boldsymbol{\eta}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) / \partial(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\boldsymbol{V}_i$ is the working variance-covariance matrix, $\boldsymbol{U}_i = (\boldsymbol{Y}_{i,1}, \boldsymbol{Y}_{i,2}, \boldsymbol{Y}_i)'$ and $\boldsymbol{\eta}_i = E(\boldsymbol{U}_i | \boldsymbol{\beta}, \boldsymbol{\gamma})$.

7.3 Dependent agreement indexes

7.3.1 Weighted least-squares approach

The weighted least-squares approach (WLS) is an extension of the GSK (Grizzle, Starmer and Koch) methodology, developed by Grizzle et al. (1969) for comparing correlated categorical data. The method was initially developed by Koch et al. (1977) and was adapted to the special case of kappa coefficients by Barnhart and Williamson (2002).

7.3.1.1 Comparison of two kappa coefficients

The weighted least-squares approach allows the comparison of several correlated agreement coefficients between two raters. Suppose that two raters classify each of N items on a K-categorical scale with two methods and let Y_{11} and Y_{12} represent the classification of the two raters with the first method and Y_{21} and Y_{22} with the second one. Suppose that interest is to determine whether the reproducibility between the two classifications differs from method to method. Because the four classifications are assessed on the same set of items, the two agreement indexes

are generally correlated and this correlation must be taken into account for valid inference.

Cross-classifying Y_{11}, Y_{12}, Y_{21} and Y_{22}, a $K \times K \times K \times K$ contingency table with cell counts y_{ijlm} $(i, j, l, m = 1, \cdots, K)$ is obtained. Interest is in two agreement indexes $\hat{\kappa}_1$ and $\hat{\kappa}_2$ obtained from the two bivariate marginal tables, i.e., the contingency tables $Y_{11} \times Y_{12}$ and $Y_{21} \times Y_{22}$ with cell counts $y_{ij..}$ and $y_{..lm}$. Note that $\hat{\kappa}_1$ is the agreement between the two readings using the first method and $\hat{\kappa}_2$ using the second one. Interest is in testing

$$H_0 : \ \kappa_1 = \kappa_2 \text{ versus } H_1 : \ \kappa_1 \neq \kappa_2 \tag{7.18}$$

and estimating the common value of the kappa coefficient if H_0 is not rejected.

Let $\boldsymbol{\pi} = (\pi_{1111}, \pi_{1112}, \cdots, \pi_{KK11}, \cdots, \pi_{KKKK})'$ denote the $K^4 \times 1$ vector of cell probabilities for the $Y_{11} \times Y_{12} \times Y_{21} \times Y_{22}$ contingency table, where $\pi_{ijlm} = P(Y_{11} = i, Y_{12} = j, Y_{21} = l, Y_{22} = m)$. Denoting $\boldsymbol{\kappa} = (\kappa_1, \kappa_2)'$, $\boldsymbol{\kappa}$ can be written as an explicit function of $\boldsymbol{\pi}$, called the *response function*, in the following form (Barnhart and Williamson, 2002):

$$\boldsymbol{\kappa} = \boldsymbol{F}(\boldsymbol{\pi}) \equiv \exp \boldsymbol{A}_4 \ln \boldsymbol{A}_3 \exp \boldsymbol{A}_2 \ln \boldsymbol{A}_1 \boldsymbol{A}_0 \boldsymbol{\pi} \tag{7.19}$$

where the matrices $\boldsymbol{A}_i$ $(i = 0, \cdots, 4)$ are defined later, depending on which kappa is used. The notation $\ln \boldsymbol{AB}$ stands for taking the napierian logarithm of each element of the matrix resulting from the multiplication of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

The weighted least-squares estimator of $\boldsymbol{\kappa}$ is

$$\widehat{\boldsymbol{\kappa}} = \exp \boldsymbol{A}_4 \ln \boldsymbol{A}_3 \exp \boldsymbol{A}_2 \ln \boldsymbol{A}_1 \boldsymbol{A}_0 \boldsymbol{P} \tag{7.20}$$

where $\boldsymbol{P}$ is the vector of the cell proportions of the K^4 table, which estimates $\boldsymbol{\pi}$. The estimated variance-covariance matrix of $\widehat{\boldsymbol{\kappa}}$ is

$$cov(\hat{\boldsymbol{\kappa}}) = \left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{P}}\right) \boldsymbol{V} \left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{P}}\right)' \tag{7.21}$$

where $\boldsymbol{V} = (diag(\boldsymbol{P}) - \boldsymbol{P}\boldsymbol{P}')/N$ is the estimated variance-covariance matrix of $\boldsymbol{P}$, $diag(\boldsymbol{P})$ denotes the diagonal matrix with $\boldsymbol{P}$ in the diagonal entry, $\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{P}}$ is the partial derivative of $\boldsymbol{F}$ with respect to $\boldsymbol{\pi}$ evaluated at $\boldsymbol{\pi} = \boldsymbol{P}$ and N is the total number of items. The partial derivative of the $\boldsymbol{F}$ function defined in Equation 7.19 has the following form

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{P}} = diag(\boldsymbol{B}_4)\boldsymbol{A}_4 diag(\boldsymbol{B}_3)^{-1}\boldsymbol{A}_3 diag(\boldsymbol{B}_2)\boldsymbol{A}_2 diag(\boldsymbol{B}_1)^{-1}\boldsymbol{A}_1\boldsymbol{A}_0 \tag{7.22}$$

where $\boldsymbol{B}_1 = \boldsymbol{A}_1\boldsymbol{A}_0\boldsymbol{P}$, $\boldsymbol{B}_2 = exp\boldsymbol{A}_2 log\boldsymbol{B}_1$, $\boldsymbol{B}_3 = \boldsymbol{A}_3\boldsymbol{B}_2$ and $\boldsymbol{B}_4 = exp\boldsymbol{A}_4 log\boldsymbol{B}_3$.

Using formulas 7.20 and 7.21, Barnhart and Williamson (2002) constructed a Wald test to test the hypothesis 7.18 using the Z-score

$$Z = \frac{\widehat{\kappa_1} - \widehat{\kappa_2}}{[var(\widehat{\kappa_1}) + var(\widehat{\kappa_2}) - 2cov(\hat{\kappa}_1, \hat{\kappa}_2)]^{\frac{1}{2}}}. \tag{7.23}$$

Barnhart and Williamson (2002) expressed Cohen's kappa coefficient as follows:

$$\begin{aligned}
\kappa &= \frac{(\pi_{11} + \pi_{22}) - (\pi_{1.}\pi_{.1} + \pi_{2.}\pi_{.2})}{1 - (\pi_{1.}\pi_{.1} + \pi_{2.}\pi_{.2})} \\
&= \exp(1 \ -1) \ln \begin{pmatrix} -1 & -1 & 1 & 0 \\ -1 & -1 & 0 & 1 \end{pmatrix} \exp \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\
&\times \ln \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{pmatrix} \\
&= \exp \boldsymbol{A}_4 \ln \boldsymbol{A}_3 \exp \boldsymbol{A}_2 \ln \boldsymbol{A}_1 \boldsymbol{\pi}.
\end{aligned} \tag{7.24}$$

The matrix $\boldsymbol{A}_1$ produces a vector with the row marginals, column marginals, diagonal sum and total sum of cell probabilities, $\boldsymbol{A}_2$ produces a vector with four main quantities in the log scale of κ, $\boldsymbol{A}_3$ produces the vector of the numerator and the denominator of κ and $\boldsymbol{A}_4$ divides the numerator by the denominator to produce κ.

The formula 7.24 was only for a single Cohen's kappa coefficient. Since two kappa coefficients have to be estimated from $\boldsymbol{\pi}$, Barnhart and Williamson (2002) presented the following formula:

$$\begin{aligned}
\begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} &= \boldsymbol{F}(\boldsymbol{\pi}) = \exp(\boldsymbol{A}_4) \ln(\boldsymbol{A}_3) \exp(\boldsymbol{A}_2) \ln(\boldsymbol{A}_1) \boldsymbol{A}_0 \boldsymbol{\pi} \\
&= \exp \begin{pmatrix} \boldsymbol{A}_{44} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{44} \end{pmatrix} \ln \begin{pmatrix} \boldsymbol{A}_{33} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{33} \end{pmatrix} \\
&\times \exp \begin{pmatrix} \boldsymbol{A}_{22} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{pmatrix} \ln \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{11} \end{pmatrix} \boldsymbol{A}_0 \boldsymbol{\pi}
\end{aligned} \tag{7.25}$$

where $\boldsymbol{A}_0$ is a $2K^2 \times K^4$ matrix with the form

$$\boldsymbol{A}_0 = \begin{pmatrix} \boldsymbol{e}'_{K^2} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{e}'_{K^2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{e}'_{K^2} \\ \boldsymbol{I}_{K^2} & \boldsymbol{I}_{K^2} & \boldsymbol{I}_{K^2} & \boldsymbol{I}_{K^2} \end{pmatrix}$$

$\boldsymbol{e}_K$ is a $K \times 1$ vector of ones, $\boldsymbol{I}_K$ is the $K \times K$ identity matrix and $\boldsymbol{0}$ is a matrix of all zeros, with dimensions conforming to the other part of the block matrices. For the different versions of the kappa coefficient, Barnhart and Williamson (2002) proposed the following expressions for the matrices $\boldsymbol{A}_{11}, \cdots, \boldsymbol{A}_{44}$.

Cohen's kappa coefficient. Barnhart and Williamson (2002) used matrices with dimensions 1×2 for $\boldsymbol{A}_{44}$, $2 \times (K+2)$ for $\boldsymbol{A}_{33}$, $(K+2) \times (2K+2)$ for $\boldsymbol{A}_{22}$ and $(2K+2) \times K^2$ for $\boldsymbol{A}_{11}$:

$$\begin{aligned} \boldsymbol{A}_{44} &= \begin{pmatrix} 1 & -1 \end{pmatrix}, \\ \boldsymbol{A}_{33} &= \begin{pmatrix} -\boldsymbol{e}'_K & 1 & 0 \\ -\boldsymbol{e}'_K & 0 & 1 \end{pmatrix}, \end{aligned}$$

$$\boldsymbol{A}_{22} = \begin{pmatrix} \boldsymbol{I}_K & \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_2 \end{pmatrix},$$

$$\boldsymbol{A}_{11} = \begin{pmatrix} \boldsymbol{e}'_K & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{e}'_K \\ \boldsymbol{I}_K & \boldsymbol{I}_K & \cdots & \boldsymbol{I}_K \\ \boldsymbol{I}_K(1) & \boldsymbol{I}_K(2) & \cdots & \boldsymbol{I}_K(K) \\ \boldsymbol{e}'_K & \boldsymbol{e}'_K & \cdots & \boldsymbol{e}'_K \end{pmatrix}$$

where $\boldsymbol{I}_K(j)$ is the j^{th} row of the identity matrix $\boldsymbol{I}_K$.

Intraclass kappa coefficient. For the intraclass kappa coefficient, Barnhart and Williamson (2002) used the same $\boldsymbol{A}_{44}$ and $\boldsymbol{A}_{33}$ matrices as for Cohen's kappa coefficient but

$$\boldsymbol{A}_{22} = \begin{pmatrix} 2\boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_2 \end{pmatrix},$$

$$\boldsymbol{A}_{11} = \begin{pmatrix} \frac{\boldsymbol{e}'_K+\boldsymbol{I}_K(1)}{2} & \frac{\boldsymbol{I}_K(1)}{2} & \cdots & \frac{\boldsymbol{I}_K(1)}{2} \\ \frac{\boldsymbol{I}_K(2)}{2} & \frac{\boldsymbol{e}'_K+\boldsymbol{I}_K(2)}{2} & \cdots & \frac{\boldsymbol{I}_K(2)}{2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\boldsymbol{I}_K(K)}{2} & \frac{\boldsymbol{I}_K(K)}{2} & \cdots & \frac{\boldsymbol{e}'_K+\boldsymbol{I}_K(K)}{2} \\ \boldsymbol{I}_K(1) & \boldsymbol{I}_K(2) & \cdots & \boldsymbol{I}_K(K) \\ \boldsymbol{e}'_K & \boldsymbol{e}'_K & \cdots & \boldsymbol{e}'_K \end{pmatrix}$$

where $\boldsymbol{A}_{22}$ and $\boldsymbol{A}_{11}$ are $(K+2)\times(K+2)$ and $(K+2)\times K^2$ matrices, respectively.

Weighted kappa coefficient. For the weighted kappa coefficient, Barnhart and Williamson (2002) used the same $\boldsymbol{A}_{44}$ matrix as for Cohen's kappa but

$$\boldsymbol{A}_{33} = \begin{pmatrix} -\boldsymbol{w}' & 1 & 0 \\ -\boldsymbol{w}' & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{A}_{22} = \begin{pmatrix} \boldsymbol{e}_K & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{e}_K & \cdots & \boldsymbol{0} & \boldsymbol{I}_K & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{e}_K & \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_2 \end{pmatrix},$$

$$\boldsymbol{A}_{11} = \begin{pmatrix} \boldsymbol{e}'_K & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{e}'_K \\ \boldsymbol{I}_K & \boldsymbol{I}_K & \cdots & \boldsymbol{I}_K \\ & & \boldsymbol{w}' & \\ \boldsymbol{e}'_K & \boldsymbol{e}'_K & \cdots & \boldsymbol{e}'_K \end{pmatrix}$$

where $\boldsymbol{w} = (w_{11}, w_{12}, \cdots, w_{KK})$ is the $K^2 \times 1$ vector of weights. The dimensions of the $\boldsymbol{A}$ matrices are $2 \times (K^2+2)$ for $\boldsymbol{A}_{33}$, $(K^2+2)\times(2K+2)$ for $\boldsymbol{A}_{22}$ and $(2K+2)\times K^2$ for $\boldsymbol{A}_{11}$.

The number of cells of the four-way contingency table $Y_{11}\times Y_{12}\times Y_{21}\times Y_{22}$ increases rapidly as the number of categories K increases. This may result in many zero cells even if the sample size is large. For valid inference using WLS, one needs to assume that the sample response functions are normally distributed and that their estimated variance-covariance matrix is nonsingular. In estimating kappa, this assumption usually requires that most of the marginal or diagonal counts in the bivariate marginal tables $Y_{11}\times Y_{12}$ and $Y_{21}\times Y_{22}$ exceed 5. Barnhart and Williamson (2002) therefore proposed to replace zeros cell counts by $1e-20$ if frequency data are used or missing data by $1e-20$ if raw data are used to treat all zeros as sampling zeros.

7.3.1.2 Comparison of several kappa coefficients

The WLS approach for comparing correlated kappa statistics can easily be extended to include discrete covariates and to study designs with multiple methods and multiple time points. If there are G methods ($G > 2$), we have K^G contingency tables and the $\boldsymbol{A}$ matrices, except $\boldsymbol{A}_0$, will have G blocks (instead of two) with the same blocks as specified previously.

7.3.2 Generalized estimating equations

Binary scale. Klar et al. (2000) proposed the use of generalized estimating equations (GEE) to identify covariates predictive of agreement, in the case of intraclass kappa coefficients. The proposed model may include arbitrary and variable number of raters per item. Generalized estimating equations (GEE1) with a logistic link function are used in order to identify covariates associated with the marginal probabilities of classification by each rater. A second set of generalized estimating equations (GEE2), based on Fisher's Z transformation, are then used to identify covariates associated with the intraclass kappa coefficient.

Suppose that N items are being assessed by R_i independent or dependent raters ($i = 1, \cdots, N$) where R_i need not be the same for all N items. The binary ratings related to item i are summarized in the $R_i \times 1$ vector $\boldsymbol{Y}_i = (Y_{i,1}, \cdots, Y_{i,R_i})'$ where the binary random variable $Y_{i,r} = 1$ when rater r classifies item i as positive and $Y_{i,r} = 0$ otherwise. Each item has a $p_1 \times 1$ covariate vector $\boldsymbol{g}_i$ including item specific covariates and a $p_2 \times 1$ covariate vector $\boldsymbol{g}_{i,r}$ ($r = 1, \cdots, R_i$) including rater specific covariates. Let $\boldsymbol{x}'_{i,r} = (\boldsymbol{g}_i, \boldsymbol{g}_{i,r})$ and $\boldsymbol{X}_i = (\boldsymbol{x}_{i,1}, \cdots, \boldsymbol{x}_{i,R_i})'$ represent the $R_i \times p$ matrix of covariates relative to item i ($p = p_1 + p_2$). The marginal distribution of $Y_{i,r}$ is Bernoulli with $\pi_{i,r} = P(Y_{i,r} = 1 | \boldsymbol{x}'_{i,r}, \boldsymbol{\beta})$ such that

$$\ln\left(\frac{\pi_{i,r}}{1-\pi_{i,r}}\right) = \ln\left(\frac{\pi_{i,r}(\boldsymbol{\beta})}{1-\pi_{i,r}(\boldsymbol{\beta})}\right) = \boldsymbol{x}'_{i,r}\boldsymbol{\beta} \tag{7.26}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. The probabilities $\pi_{i,r}(\boldsymbol{\beta})$ can be grouped together to form a vector $\boldsymbol{\pi}_i(\boldsymbol{\beta})$ containing the marginal probabilities of success

$$\boldsymbol{\pi}_i(\boldsymbol{\beta}) = E(\boldsymbol{Y}_i | \boldsymbol{x}_{i,r}, \boldsymbol{\beta}) = (\pi_{i,1}, \cdots, \pi_{i,R_i})'.$$

Following Shoukri and Mian (1996), the intraclass coefficient of agreement between $Y_{i,s}$ and $Y_{i,t}$, $\kappa_{i,st}$, can be expressed as

$$\kappa_{i,st} = \kappa_{i,st}(\boldsymbol{\gamma}) = \frac{2[\pi_{i,st} - \pi_{i,s}\pi_{i,t}]}{\pi_{i,s}(1-\pi_{i,t}) + \pi_{i,t}(1-\pi_{i,s})} \tag{7.27}$$

where $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters, and solving Equation 7.27 for $\pi_{i,st}$, we get

$$\pi_{i,st} = \pi_{i,s}\pi_{i,t} + \kappa_{i,st}\frac{[\pi_{i,s}(1-\pi_{i,t}) + \pi_{i,t}(1-\pi_{i,s})]}{2}. \tag{7.28}$$

Klar et al. (2002) accomplished the adjustment for covariates associated with the intraclass kappa coefficient using the model

$$g(\kappa_{i,st}) = \ln\left(\frac{1+\kappa_{i,st}}{1-\kappa_{i,st}}\right) = \boldsymbol{z}'_{i,st}\boldsymbol{\gamma} \tag{7.29}$$

where $\boldsymbol{z}'_{i,st}$ is a $q \times 1$ vector of covariates. The link function $g(.)$, used to model kappa, is selected to avoid the need of constraints on the vector of parameters $\boldsymbol{\gamma}$ which would have been the case, for example, if the identity link had been used since $-1 \leq \kappa_{i,st} \leq 1$.

Klar et al. (2000) obtained parameter estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by specifying and solving two sets of estimating equations. The first set of estimating equations for $\boldsymbol{\beta}$ is given by

$$\boldsymbol{u}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \hat{\boldsymbol{D}}'_i \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}})) = \boldsymbol{0} \tag{7.30}$$

where the $p \times R_i$ matrix $\boldsymbol{D}_i = \partial\boldsymbol{\pi}_i/\partial\boldsymbol{\beta}$, $\boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is a $R_i \times R_i$ 'working' covariance matrix of $\boldsymbol{Y}_i$. The correlation does not need to be correctly specified for consistent parameter estimation, although the closer $\hat{\boldsymbol{V}}_i$ is of the true covariance matrix of $\boldsymbol{Y}_i$, the greater the efficiency of $\hat{\boldsymbol{\beta}}$. Klar et al. (2000) considered two approaches to construct estimating equations for $\boldsymbol{\gamma}$. These approaches can be distinguished by their use of unconditional (Prentice, 1988) and conditional (Carey et al., 1993) residuals.

The unconditional residuals are expressed as deviations about the unconditional expectations, i.e.,

$$ur_{i,st} = Y_{i,s}Y_{i,t} - E(Y_{i,s}Y_{i,t}|\boldsymbol{x}_{i,s}, \boldsymbol{x}_{i,t}) = U_{i,st} - \pi_{i,st} \tag{7.31}$$

where $U_{i,st} = Y_{i,s}Y_{i,t}$. Klar et al. (2000) grouped the unconditional residuals to form a $R_i(R_i-1)/2$ vector $\boldsymbol{ur} = \boldsymbol{U}_i - \boldsymbol{\lambda}_i$, where $\boldsymbol{U}_i = (U_{i,12}, U_{i,13}, \cdots, U_{i,(R_i-1)R_i})'$ and $\boldsymbol{\lambda}_i = (\pi_{i,12}, \pi_{i,13}, \cdots, \pi_{i,(R_i-1)R_i})'$. The second set of estimating equations (GEE2) for $\boldsymbol{\gamma}$ is then given by

$$\boldsymbol{u}_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{N} \hat{\boldsymbol{A}}'_i \hat{\boldsymbol{H}}_i^{-1}(\boldsymbol{U}_i - \boldsymbol{\lambda}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})) = \boldsymbol{0} \tag{7.32}$$

where the $q \times [R_i(R_i-1)/2]$ matrix $\boldsymbol{A}'_i = \partial\boldsymbol{\lambda}_i/\partial\boldsymbol{\gamma}$, and $\boldsymbol{H}_i \approx cov(\boldsymbol{U}_i)$.

Conditional residuals are expressed as deviations about the conditional expectations, i.e.,

$$cr_{i,st} = Y_{i,t} - E[Y_{i,t}|y_{i,s}, \boldsymbol{x}_{i,s}, \boldsymbol{x}_{i,t}] = U_{i,st} - \eta_{i,st} \tag{7.33}$$

where $U_{i,st} = Y_{i,t}$ and

$$\begin{aligned}\boldsymbol{\eta}_{i,st} &= E(Y_{i,t}|Y_{i,s} = y_{i,s}, \boldsymbol{x}_{i,s}, \boldsymbol{x}_{i,t}) = P(Y_{i,t} = 1|Y_{i,s} = y_{i,s}, \boldsymbol{x}_{i,s}, \boldsymbol{x}_{i,t}) \\ &= y_{i,s}\left(\frac{\pi_{i,st}}{\pi_{i,s}}\right) + (1 - y_{i,s})\left(\frac{\pi_{i,t} - \pi_{i,st}}{1 - \pi_{i,s}}\right).\end{aligned}$$

Klar et al. (2000) then grouped the conditional residuals to form a $R_i(R_i - 1)/2$ vector $\boldsymbol{cr} = \boldsymbol{U}_i - \boldsymbol{\eta}_i$, where $\boldsymbol{U}_i = (U_{i,12}, U_{i,13}, \cdots, U_{i,(R_i-1)R_i})'$ and $\boldsymbol{\eta}_i = (\eta_{i,12}, \eta_{i,13}, \cdots, \eta_{i,(R_i-1)R_i})'$. The second set of estimating equations (GEE2) for $\boldsymbol{\gamma}$ is then given by

$$\boldsymbol{u}_{\gamma}(\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{N} \hat{\boldsymbol{C}}_i' \hat{\boldsymbol{W}}_i^{-1}(\boldsymbol{U}_i - \boldsymbol{\eta}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})) = \boldsymbol{0} \tag{7.34}$$

where the $q \times [R_i(R_i - 1)/2]$ matrix $\boldsymbol{C}_i' = \partial\boldsymbol{\eta}_i/\partial\boldsymbol{\gamma}$, and $\boldsymbol{W}_i \approx diag(cov(\boldsymbol{U}_i))$.

Inferences on the intraclass kappa coefficient can be constructed using the joint distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ which is, asymptotically, multivariate Normal. Note that more precise estimators of $\boldsymbol{\gamma}$ and hence of kappa are obtained by constructing the estimating equations using conditional residuals. Lipsitz and Fitzmaurice (1996) have shown that the gain in efficiency obtained by constructing estimating equations using conditional as opposed to unconditional residuals is particularly notable when the number of raters per item is variable.

Multinomial scale. Williamson et al. (2000) proposed the use of generalized estimating equations to model dependent agreement indexes when the scale is categorical in a way similar to Klar et al. (2000). Suppose again that N items are assessed by R_i independent or dependent raters. The response of interest is a categorical outcome with K categories denoted $Y_{i,r}$, where $Y_{i,r} = k$ if response of rater r for item i falls in category k, $r = 1, \cdots, R_i$ and $k = 1, \cdots, K$. The $R_i(K-1)\times 1$ response vector $\boldsymbol{Y}_i$ consists of the binary random variables $Y_{ik,r}$, where $Y_{ik,r} = 1$ if $Y_{i,r} = k$ ($\boldsymbol{Y}_i = (Y_{i1,1}, \cdots, Y_{i(K-1),1}, \cdots, Y_{i1,(R_i-1)}, \cdots, Y_{i(K-1),(R_i-1)})'$). For ordinal responses, the marginal cumulative probabilities of response, $\nu_{ik,r} = P(Y_{i,r} \leq k)$, $k = 1, \cdots, K-1$, are modeled. Let

$$\pi_{ik,r} = \pi_{ik,r}(\boldsymbol{\beta}) = P(Y_{i,r} = k) = P(Y_{ik,r} = 1) = \nu_{ik,r} - \nu_{i(k-1),r}$$

denote the marginal probabilities. These probabilities can be grouped to form a $R_i(K-1) \times 1$ vector $\boldsymbol{\pi}_i$. The vectors $\boldsymbol{Y}_i$ and $\boldsymbol{\pi}_i$ require only $R_i(K-1)$ elements instead of R_iK since $\sum_{k=1}^{K} Y_{ik,r} = 1$, for $i = 1, \cdots, N$ and $r = 1, \cdots, R_i$. Let $\boldsymbol{x}_{ik,r}$ be the $(p + K - 1) \times 1$ vector of covariates relative to item i and rater r, which consists of covariates for the $K - 1$ cutpoints of the categorical response and p covariates relative to the raters and the items. Williamson et al. (2000) related

the cumulative marginal response probabilities to the covariates through the link function $g(.)$ and the $(p + K - 1) \times 1$ marginal parameter vector $\boldsymbol{\beta}$,

$$g(\nu_{ik,r}) = \boldsymbol{x}'_{ik,r}\boldsymbol{\beta}. \tag{7.35}$$

For example, $g(.)$ might be the cumulative logit function resulting in a proportional odds model for ordinal responses or the polytomous link function for nominal responses.

Williamson et al. (2000) then consider the joint distribution of raters s and t for item i $(Y_{i,s}Y_{i,t})$ with

$$\pi_{ijk,st} = P(Y_{i,s} = j, Y_{i,t} = k|\boldsymbol{x}_{i,s}, \boldsymbol{x}_{i,t}) \tag{7.36}$$

denoting the associated probabilities $(j, k = 1, \cdots, K)$ and then defined the agreement index

$$\kappa_{i,st} = \frac{P_{oi,st} - P_{ei,st}}{1 - P_{ei,st}} \tag{7.37}$$

with

$$P_{oi,st} = \sum_{j=1}^{K} \pi_{ijj,st} \tag{7.38}$$

and

$$P_{ei,st} = \sum_{j=1}^{K} \pi_{ij,s}\pi_{ij,t}. \tag{7.39}$$

Williamson et al. (2000) next consider the regression model for Cohen's kappa coefficient, following Klar et al. (2000) to avoid restrictions on the parameter space (see Equation 7.29). The first set of generalized estimating equations (GEE1) relative to the marginal distribution of the responses is then defined by

$$\boldsymbol{u}_{\beta}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \hat{\boldsymbol{D}}'_i \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}})) = \mathbf{0} \tag{7.40}$$

where the $(p + K - 1) \times R_i(K - 1)$ matrix $\boldsymbol{D}_i = \partial\boldsymbol{\pi}_i/\partial\boldsymbol{\beta}$ and $\boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is a $R_i(K-1) \times R_i(K-1)$ 'working' covariance matrix of $\boldsymbol{Y}_i$. Williamson et al. (2000) proposed a second set of estimating equations (GEE2) for the joint distribution of responses by noting that

$$P_{oi,st} = P_{ei,st} + \kappa_{i,st}(1 - P_{ei,st}). \tag{7.41}$$

Following Liang et al. (1992), they considered a product of indicator variables. Let

$$U_{i,st} = \sum_{k=1}^{K} Y_{ik,s}Y_{ik,t} \tag{7.42}$$

be a binary random variable depicting agreement between raters s and t with $E(U_{i,st}) = P_{oi,st}$. When considering R_i raters, they are $R_i(R_i-1)/2$ distinct pairs. Hence, $\boldsymbol{U}_i' = (U_{i,12}, U_{i,13}, \cdots, U_{i,(R_i-1)R_i})$ and $\boldsymbol{P}_{oi}' = (P_{oi,12}, P_{oi,13}, \cdots, P_{oi,(R_i-1)R_i})$ are vectors of dimension $R_i(R_i-1)/2 \times 1$ with $E(\boldsymbol{U}_i) = \boldsymbol{P}_{oi}$. The kappa coefficient is then estimated by solving a second set of estimating equations (GEE2)

$$\boldsymbol{u}_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{N} \hat{\boldsymbol{C}}_i' \hat{\boldsymbol{W}}_i^{-1} (\boldsymbol{U}_i - \boldsymbol{P}_{oi}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})) = \boldsymbol{0} \tag{7.43}$$

where the $q \times [R_i(R_i-1)/2]$ matrix $\boldsymbol{C}_i' = \partial \boldsymbol{P}_{oi}/\partial \boldsymbol{\gamma}$ and $\boldsymbol{W}_i \approx diag(cov(\boldsymbol{U}_i))$.

Williamson et al. (2000) used Fisher scoring algorithm to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(m+1)} &= \hat{\boldsymbol{\beta}}^{(m)} + \left[\sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{D}}_i\right]^{-1} \left[\sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \hat{\boldsymbol{V}}_i^{-1} \left\{\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}^{(m)})\right\}\right], \\
\hat{\boldsymbol{\gamma}}^{(m+1)} &= \hat{\boldsymbol{\gamma}}^{(m)} + \left[\sum_{i=1}^{N} \hat{\boldsymbol{C}}_i' \hat{\boldsymbol{W}}_i^{-1} \hat{\boldsymbol{C}}_i\right]^{-1} \left[\sum_{i=1}^{N} \hat{\boldsymbol{C}}_i' \hat{\boldsymbol{W}}_i^{-1} \left\{\boldsymbol{U}_i - \boldsymbol{P}_{oi}(\hat{\boldsymbol{\gamma}}^{(m)}, \hat{\boldsymbol{\beta}})\right\}\right].
\end{aligned}$$

Finally, they used Liang and Zeger (1986) empirically corrected variance estimate of $\hat{\boldsymbol{\beta}}$, i.e.,

$$V_\beta = \left[\sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{D}}_i\right]^{-1} \left[\sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \hat{\boldsymbol{V}}_i^{-1} cov(\boldsymbol{Y}_i) \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{D}}_i\right] \left[\sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{D}}_i\right]^{-1} \tag{7.44}$$

and Prentice (1988) empirically corrected variance estimate of $\hat{\boldsymbol{\gamma}}$

$$V_\gamma = \boldsymbol{B}\boldsymbol{\Lambda}_{11}\boldsymbol{B}' + \boldsymbol{B}\boldsymbol{\Lambda}_{12}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{\Lambda}_{21}\boldsymbol{B}' + \boldsymbol{E}\boldsymbol{\Lambda}_{22}\boldsymbol{E} \tag{7.45}$$

where

$$\begin{aligned}
\boldsymbol{B} &= \left[\sum_{i=1}^{N} \hat{\boldsymbol{C}}_i' \hat{\boldsymbol{W}}_i^{-1} \hat{\boldsymbol{C}}_i\right]^{-1} \left[\sum_{i=1}^{N} \hat{\boldsymbol{C}}_i' \hat{\boldsymbol{W}}_i^{-1} \frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\beta}}\right] \left[\sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{D}}_i\right]^{-1} \\
\boldsymbol{E} &= \left[\sum_{i=1}^{N} \hat{\boldsymbol{C}}_i' \hat{\boldsymbol{W}}_i^{-1} \hat{\boldsymbol{C}}_i\right]^{-1} \\
\boldsymbol{\Lambda}_{11} &= \sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \hat{\boldsymbol{V}}_i^{-1} cov(\boldsymbol{Y}_i) \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{D}}_i \\
\boldsymbol{\Lambda}_{12} &= \sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \hat{\boldsymbol{V}}_i^{-1} cov(\boldsymbol{Y}_i, \boldsymbol{U}_i) \hat{\boldsymbol{W}}_i^{-1} \hat{\boldsymbol{D}}_i \\
\boldsymbol{\Lambda}_{21} &= \boldsymbol{\Lambda}_{12}' \\
\boldsymbol{\Lambda}_{22} &= \sum_{i=1}^{N} \hat{\boldsymbol{C}}_i' \hat{\boldsymbol{W}}_i^{-1} cov(\boldsymbol{U}_i) \hat{\boldsymbol{W}}_i^{-1} \hat{\boldsymbol{C}}_i
\end{aligned}$$

Williamson et al. (2000) conducted analyses using simulated data to assess the performance of their method. They examined the effects of a misspecified marginal distribution on the association parameters and determined how well the empirically corrected standard error estimate performs with small to moderate sample size. It results that the standard error estimate of the kappa coefficient may be biased in small samples ($N \leq 30$). The association model needs not to be correctly specified for unbiased estimation with the marginal model because the first set of estimating equations does not involve kappa. But, it is crucial to model the marginal distribution carefully even when interest is only in the agreement between responses because omission of an important (significant) marginal covariate may produce a biased estimate of the kappa coefficient.

Ordinal scale. Gonin et al. (2000) proposed the use of generalized estimating equations to model dependent agreement data, using the weighted kappa coefficient. They used an ordinal logistic regression model to identify covariates that are associated with the marginal probability of classification by each rater and a second set of estimating equations, based on the Fisher's Z transformation, to model the weighted kappa coefficient as a function of covariates. Using the same notation as in previous section, Gonin et al. (2000) considered the weighted agreement index

$$\kappa_{wi,st} = \frac{P_{owi,st} - P_{ewi,st}}{1 - P_{ewi,st}} \tag{7.46}$$

with

$$P_{owi,st} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} \pi_{ijk,st} \tag{7.47}$$

and

$$P_{ewi,st} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} \pi_{ij,s} \pi_{ik,t} \tag{7.48}$$

where the weights w_{jk} satisfy $w_{jj} = 1$ and $0 \leq w_{jk} \leq 1$. Next, they considered the regression model for the weighted kappa coefficient similarly to Klar et al. (2000) (see Equation 7.29). The first set of generalized estimating equations (GEE1) to estimate the vector of parameters $\boldsymbol{\beta}$ used by Gonin et al. (2000) is defined by Equation 7.40.

Denote by $U_{wi,st} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} Y_{ij,s} Y_{ik,t}$, the product of indicator variables. To define the second set of estimating equations, Gonin et al. (2000) noted that

$$E\left(U_{wi,st} | \boldsymbol{X}_i\right) = P_{owi,st} \tag{7.49}$$

by assuming that rater specific covariates from rater t does not influence the ratings

of rater $s, t \neq s$, i.e. $E[Y_{ij,s}Y_{ik,t}|\boldsymbol{X}_i] = E[Y_{ij,s}Y_{ik,t}|\boldsymbol{x}_{i,s}, \boldsymbol{x}_{i,t}]$. Moreover,

$$P_{owi,st} = \kappa_{wi,st}\left(1 - \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}\pi_{ij,s}\pi_{ik,t}\right) + \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}\pi_{ij,s}\pi_{ik,t}, \tag{7.50}$$

implying that $U_{wi,st} - P_{owi,st}$ is unbiased for $\mathbf{0}$. They therefore proposed, for GEE2,

$$\boldsymbol{u}_\gamma(\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{N}\sum_{s<t} \hat{\boldsymbol{C}}'_{wi,st}\hat{W}^{-1}_{wi,st}\left(U_{wi,st} - P_{owi,st}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\right) = \mathbf{0} \tag{7.51}$$

where $\hat{\boldsymbol{C}}'_{wi,st} = \partial P_{owi,st}/\partial\boldsymbol{\gamma}$, W is the 'working' variance of $U_{wi,st}$, $W \sim var(U_{wi,st})$.

Using Taylor series expansions similarly to Prentice (1988), and assuming that the model for $\kappa_{wi,st}$, $\pi_{ij,s}$ and $\pi_{ik,t}$ are correctly specified, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is consistent and asymptotically Normal, with variance-covariance matrix $\boldsymbol{V}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$. This covariance matrix can be consistently estimated using a Jackknife estimate (Lipsitz et al., 1990):

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}') = \sum_{i=1}^{N}((\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i} - (\bar{\boldsymbol{\beta}}', \bar{\boldsymbol{\gamma}}'))'((\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i} - (\bar{\boldsymbol{\beta}}', \bar{\boldsymbol{\gamma}}')) \tag{7.52}$$

where $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i}$ is the estimate of $(\boldsymbol{\beta}', \boldsymbol{\gamma}')$ obtained by deleting the R_i ratings on item i and $(\bar{\boldsymbol{\beta}}', \bar{\boldsymbol{\gamma}}')$ is the average of the $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')_{-i}$ over the N individuals. Gonin et al. (2000) proposed then to solve the two sets of estimating equations using a Fisher scoring type of algorithm and an exchangeable working correlation matrix for $\boldsymbol{Y}_i$. They shown that this permits to completely specify the working correlation of $U_{wi,st}$.

7.4 Simulations

The bootstrap method of Vanbelle and Albert (2008) was applied to simulated data sets in order to study the behavior of the type I error (α) of the homogeneity test for $G = 3$ Cohen's kappa coefficient and compare the results with the WLS and GEE2 approaches. Each simulation consisted in applying the bootstrap method to 3000 data sets generated under the null hypothesis $H_0 : \kappa_1 = \kappa_2 = \kappa_3$ and to determine the number of times H_0 was rejected. The simulated data set was based on 4 binary random variables X, Y, Z and V. The agreement between X and Y (κ_{XY}), X and Z (κ_{XZ}) and X and V (κ_{XV}) were compared using the bootstrap method with $q = 2000$ iterations. Simulations were repeated for 3 sample sizes (50, 75 and 100) and 5 levels of agreement (κ=0, 0.2, 0.4, 0.6 and 0.8).

To obtain a given level of agreement (κ), 2 vectors of size N from binary random variables (U and W) were generated. Then, a vector of size N with uniform random numbers between 0 and 1 was generated. Each time the random uniform number was less than or equal to the given level of agreement (κ), the value of W was changed into the value of U, otherwise it remained unchanged. The kappa coefficient was derived from the 2×2 table obtained by cross-classifying the vectors U and W. The codes for the simulations were written in R language using uniform random number generator with seed equal to 2. The GEE2 (Williamson et al., 2000) and the WLS (Barnhart and Williamson, 2002) approaches were also applied to the 3000 simulated data sets. Results are summarized in Table 7.1.

Table 7.1. Type I error for the comparison of $G = 3$ correlated kappa coefficients, according to κ level and sample size (figures are based on 3000 simulations each)

		κ level				
Sample size	Method	0	0.2	0.4	0.6	0.8
50	Bootstrap[a]	0.065	0.069	0.061	0.076	0.056
	GEE2	0.067	0.061	0.063	0.052	0.044
	WLS	0.0027	0.037	0.062	0.0769	0.064
75	Bootstrap	0.070	0.061	0.061	0.063	0.063
	GEE2	0.046	0.058	0.057	0.051	0.040
	WLS	0.0030	0.040	0.060	0.071	0.069
100	Bootstrap	0.089	0.065	0.064	0.061	0.058
	GEE2	0.057	0.054	0.050	0.053	0.040
	WLS	0.0027	0.037	0.055	0.064	0.064

[a] $q = 2000$

It is seen that type I error rates obtained with the bootstrap method are slightly but systematically higher than the expected 5% nominal level. While the GEE2 approach appears to be optimal, the bootstrap was better than the WLS, at least for elevated κ values. However, the bootstrap method may be preferred to the GEE2 approach because of the ease of implementation in all settings as compared to the GEE2 method, which requires the writing of a lengthy and specific program for each particular problem.

7.5 Examples

7.5.1 Blood clots detection

The presence of blood clots was assessed on 50 patients (23 women and 27 men) with a reference method (Standard) and two new methods (Method 1 and Method 2) by two medical raters (see Chapter 2, Section 2.6.2). The aim of the study was to compare the agreement obtained between Method 1 and the Standard method to the agreement obtained between Method 2 and the Standard.

Weighted least-squares approach. Cohen's kappa coefficient was 0.41 ± 0.12 between Method 1 and the Standard method and 0.71 ± 0.10 between Method 2 and the Standard. A test of equality of these agreement indexes using the WLS approach resulted in a chi-square value of 6.16 ($p = 0.013$). A better agreement with the Standard method was observed for Method 2 than for Method 1. Therefore, Method 2 should be preferred to Method 1 (this is consistent with the results of the bootstrap method (Vanbelle and Albert, 2008)). The estimated correlation between the two kappa estimates was 0.43.

When testing the effect of sex on Cohen's kappa coefficients, we obtained an estimate of -0.23 ($p = 0.20$), resulting in no agreement difference between men and women. For Method 1, the estimate of Cohen's kappa coefficients were 0.26 and 0.49 for men and women, respectively. For Method 2, the estimate of Cohen's kappa coefficients were 0.58 and 0.81 for men and women, respectively.

GEE2 and initial method of Lipsitz et al. (2001). Although constraints are needed to ensure that Cohen's kappa coefficient is comprised in the interval $[-1, 1]$, an identity link was used to simplify the interpretation of the parameters and permit the comparison of the GEE2 and the method proposed by Lipsitz et al. (2001). Note that this is not statistically correct to apply the alternative method to test for equality between the agreement obtained for Method 1 and Method 2 since the data are dependent. The model for Cohen's kappa coefficient was

$$\begin{aligned}\kappa &= Intercept + \beta_1 SEX + \beta_2 AGE + \beta_3 METHOD1 + \beta_4 AGE \times METHOD1 \\ &\quad + \beta_5 SEX \times METHOD1 + \beta_6 AGE \times SEX \end{aligned} \tag{7.53}$$

where $METHOD1 = 1$ when agreement is searched between Method 1 and the Standard method ($METHOD1 = 0$ otherwise), $SEX = 1$ for men ($SEX = 0$ for women) and AGE is patient's age (years). The resulting parameter estimates are displayed in Table 7.2 for the GEE2 and the method of Lipsitz et al. (2001).

Method 1 showed lower agreement with the Standard method than with Method 2 ($p = 0.024$ with GEE2, $p = 0.034$ with the method of Lipsitz et al. (2001)).

Table 7.2. Estimates of the model of Cohen's kappa coefficient for the blood clots detection example with GEE2 and method of Lipsitz et al. (2001)

	GEE2			Lipsitz et al. (2001)		
Parameter	$\hat{\beta}$	SE	p-value	$\hat{\beta}$	SE	p-value
$Intercept$	1.42	0.39	< 0.0001	1.20	0.24	< 0.0001
SEX	-1.38	0.67	0.038	-1.15	0.56	0.040
AGE	-0.0083	0.0065	0.20	-0.0060	0.0052	0.25
$METHOD1$	-1.25	0.55	0.024	-1.32	0.63	0.034
$AGE \times METHOD1$	0.013	0.0076	0.093	0.014	0.0081	0.086
$SEX \times METHOD1$	0.14	0.26	0.53	0.16	0.28	0.56
$AGE \times SEX$	0.017	0.0095	0.073	0.014	0.0077	0.076

Agreement was also lower for men than for women ($p = 0.038$ with GEE2, $p = 0.040$ with the alternative method). There was no evidence for an effect of age. Thus, Method 2 should be preferred to Method 1. This result is consistent with the WLS and the bootstrap approaches. Note that the parameter estimates and the standard errors were similar for both GEE2 and the method of Lipsitz et al. (2001).

Method of Lipsitz et al. (2001). The alternative procedure proposed by Lipsitz et al. (2001) was applied to study the effect of patient's age and sex on the kappa coefficients obtained for Method 1 and Method 2, separately. The marginal model relative to the classification of the patients for each method was

$$\text{logit}P(Y_{i,r} = 1) = Intercept + \beta_1 SEX + \beta_2 AGE, \quad (r = 1, 2; i = 1, \cdots, N) \quad (7.54)$$

where $SEX = 1$ for men and $SEX = 0$ for women and AGE is patient's age (years). The parameter estimates are given in Table 7.3.

Table 7.3. Parameter estimates of the marginal models relative to each method for the blood clots detection example (p for pvalue)

	Method 1			Method 2			Standard method		
Parameter	$\hat{\beta}$	SE	p	$\hat{\beta}$	SE	p	$\hat{\beta}$	SE	p
$Intercept$	1.48	1.52	0.33	0.57	1.45	0.15	1.24	1.45	0.39
SEX	-1.83	0.66	0.0053	-0.99	0.61	0.10	-0.89	0.61	0.14
AGE	-0.0028	0.020	0.89	-0.0066	0.0192	0.73	-0.016	0.019	0.40

The marginal distribution obtained with Method 1 was related to patient's sex. The probability to detect blood clots was higher in women than in men with

Method 1 ($p < 0.01$). This was not the case for Method 2 and for the Standard method. There was no effect of age.

For both methods, the effect of sex and patient's age was tested on Cohen's kappa coefficient obtained by the Standard method with the model

$$\kappa_g = Intercept + \beta_1 SEX + \beta_2 AGE \tag{7.55}$$

where $g = 1$ for Method 1 and $g = 2$ for Method 2. The parameter estimates are displayed in Table 7.4.

Table 7.4. Parameter estimates of the model for the kappa coefficients obtained with Method 1 and Method 2 for the blood clots detection example

	Method 1			Method 2		
Parameter	$\hat{\beta}$	SE	p-value	$\hat{\beta}$	SE	p-value
Intercept	-1.20	0.46	0.0093	1.13	0.26	<0.0001
SEX	-0.049	0.22	0.83	-0.34	0.21	0.11
AGE	0.023	0.0056	<0.0001	-0.0038	0.0039	0.34

No covariate effect was found for the agreement obtained for Method 2. On the other hand, the agreement between Method 1 and the Standard method increased with patient's age. When computing Cohen's kappa coefficient for patients below median age (73.5 years) and above median age, we obtained Table 7.5, confirming the effect of age on the agreement obtained with Method 1. Note that results were consistent with the results obtained with the method introduced by Fleiss (1981) (see Chapter 6, Section 6.5.1), where there was no sex effect on the agreement.

Table 7.5. Cohen's kappa coefficient $\pm SE$ for each method according to patient's median age for the blood clots detection example

Age	N	Method 1	Method 2
≤73.5 years	27	0.080 ± 0.18	0.63 ± 0.17
>73.5 years	23	0.76 ± 0.13	0.76 ± 0.13

7.5.2 Cervical ectopy size

Cervical ectopy size of 85 women was determined on a 4-category scale by two medical raters with direct visual assessment and the computerized planimetry

method. To test if agreement between the two raters was the same for the planimetry and the visual methods, the WLS, GEE2 and the method proposed by Lipsitz et al. (2001) (although data are repeated) were used to test the equality of the agreement obtained between the two raters with the visual and the planimetry methods. Since the alternative method is only designed for unweighted kappa coefficients, Cohen's kappa coefficient was used as measure of agreement to permit the comparison of the different approaches. Results are summarized in Table 7.6 with the estimated correlation between the agreement indexes for the visual and the planimetry method.

Table 7.6. Results of the WLS, GEE2 and alternative method when testing for equality between the agreement for the visual and the planimetry methods in the cervical ectopy size example

	Visual assessment	Planimetry method		
Method	$\hat{\kappa}_v \pm SE$	$\hat{\kappa}_p \pm SE$	p-value	$corr(\hat{\kappa}_v, \hat{\kappa}_p)$
WLS	0.34 ± 0.068	0.63 ± 0.067	0.0022	0.064
GEE2	0.33 ± 0.099	0.63 ± 0.041	0.0019	0.0041
ALTERNATIVE	0.33 ± 0.070	0.63 ± 0.068	0.0026	-0.020

The three approaches lead to the same conclusion, the agreement between the two raters was higher with the planimetry method than with direct visual assessment. The estimated correlation between the two agreement indexes was negligible.

GEE2. The marginal probabilities of classification by each rater were modeled following

$$\text{logit}P(Y_i \leq k) = \alpha_k + \beta_1 PL + \beta_2 R1 + \beta_3 PL \times R1 \ (k = 1, 2, 3; i = 1, \cdots, N) \quad (7.56)$$

where $R1 = 1$ for rater 1 ($R1 = 0$ for rater 2) and $PL = 1$ for the planimetry method ($PL = 0$ for the visual method). The resulting parameter estimates are displayed in Table 7.7.

The probability of being classified with smaller ectopy sizes was lower for rater 1 than for rater 2 ($p = 0.0009$) and higher with the planimetry method than with visual assessment ($p < 0.0001$). The estimates of Cohen's kappa coefficients were given in Table 7.6.

Table 7.7. Parameter estimates of the GEE2 model in the cervical ectopy size example

Parameter	$\hat{\beta}$	SE	p-value
α_1	-0.77	0.19	<0.0001
α_2	0.86	0.20	<0.0001
α_3	1.65	0.25	<0.0001
PL	0.60	0.18	0.0009
$R1$	-0.79	0.16	<0.0001
$PL \times R1$	0.55	0.20	0.007

Method of Lipsitz et al. (2001). The marginal probabilities were modeled with respect to the method used for each rater

$$\text{logit}P(Y_{i,r} \leq k) = \alpha_{kr} + \beta_1 PL, \quad (r = 1, 2; k = 1, 2, 3; i = 1, \cdots, N). \tag{7.57}$$

Results are given in Table 7.8. The marginal probabilities distribution of both raters was related to the method used. The probability to be classified has having smaller ectopy sizes was higher with the planimetry method than with the visual assessment.

Table 7.8. Parameter estimates of the marginal models for each rater obtained with the alternative approach corresponding to the classification of cervical ectopy sizes

	Rater 1			Rater 2		
Parameter	$\hat{\beta}$	SE	p-value	$\hat{\beta}$	SE	p-value
α_1	-1.62	0.25	<0.0001	-0.80	0.22	0.0003
α_2	0.10	0.21	0.62	0.74	0.22	0.0006
α_3	0.73	0.22	0.0009	1.75	0.27	<0.0001
PL	1.15	0.29	<0.0001	0.60	0.28	0.034

The model for Cohen's kappa coefficient was

$$\kappa = Intercept + \beta PL. \tag{7.58}$$

The parameter estimates relative to this model are displayed in Table 7.9. Cohen's kappa coefficient was better for the planimetry method than for the visual assessment ($p = 0.0030$)

Table 7.9. Parameter estimates of the model for the Cohen's kappa coefficient using the alternative method for the cervicl ectopy size example

Parameter	$\hat{\beta}$	SE	p-value
Intercept	0.33	0.070	<0.0001
PL	0.30	0.098	0.0030

7.6 Discussion

Only some of the existing modeling techniques were reviewed in this chapter. Firstly, methods allowing the modelization of independent agreement indexes were reviewed (Lipsitz et al., 2001, 2003). These methods are heuristic but permit to model directly the kappa coefficient in function of continuous and categorical covariates. Nevertheless, they do not permit the modeling of weighted agreement indexes and do not allow to study correlated agreement indexes. However, one can possibly use generalized estimating equations instead of simple logistic regression to allow for dependent agreement indexes. Indeed, when modifying their approach, Lipsitz et al. (2003) have shown that their method was equivalent to using one set of generalized estimating equations. The generalization of the method of Lipsitz et al. (2001, 2003) to weighted agreement indexes might be an interesting theme for future research.

Then, approaches allowing for dependent agreement indexes were reviewed. The weighted least-squares approach permits the comparison of several correlated agreement coefficients between two raters. This approach has to be viewed more as a comparison method than a modeling technique since it gives estimated values of the agreement coefficients rather than estimates of the effect of covariates, i.e model kappa coefficients with respect to covariates. The method is easy to implement using, for example, PROC CATMOD in the SAS software but is, however, restricted to categorical covariates. On the other hand, the GEE2 approach is based on the estimation of two generalized estimating equations, one to characterize the marginal classifications of the raters and a second to study the effect of covariates on the agreement index obtained between the raters. The GEE2 approach has the advantage of allowing for continuous covariates but requires programming work since there is no procedure for GEE2 in standard statistical packages, to our knowledge. Moreover, efficiency of the estimates depends on the choice of a working correlation matrix although consistency does not. For categorical covariates, Miller et al. (1993) have shown that the WLS approach for analysing multi-way contingency tables is asymptotically equivalent to the GEE2 approach under a

common unspecified working correlation matrix for all covariate patterns.

The WLS and GEE2 approaches were compared to the bootstrap method of Vanbelle and Albert (2008). The WLS method developed by Barnhart and Williamson (2002) and the GEE2 approach of Williamson et al. (2000) led to the same conclusions as the bootstrap procedure for both examples, although estimates of the kappa coefficients obtained with the bootstrap method were slightly biased. However, Efron and Tibshirani (1993) suggested that if the estimate of the bias ($\hat{bias}$) is small compared to the estimate of the standard error ($\hat{SE}$), i.e. $\hat{bias}/\hat{SE} \leq 0.25$, the bias can be ignored. Otherwise, it may be an indication that $\hat{\kappa}$ is not an appropriate estimate of the parameter κ. The bootstrap approach also yields slightly higher standard errors than the WLS and the GEE2 methods, as it was expected from the results of the simulations. Indeed, the type I errors obtained with the bootstrap method were more liberal than those with the GEE2 method, in particular if the sample size (N) was small with respect to the number (G) of kappas to be compared. This finding confirms the remark made by McKenzie et al. (1996). Nevertheless, the type I error obtained by the bootstrap remains acceptable although it is recommended to use more than 1000 bootstrap iterations when the number of agreement coefficients to be compared is greater than 2. The bootstrap method outlined in Section 6.4.3 can be easily implemented in many statistical packages and programming languages since the method merely requires the generation of random uniform numbers and simple matrix calculations. By contrast, modeling techniques require specific programming for each problem encountered in practice. Their use is nevertheless highly recommended when it comes to account for many covariates. Lin et al. (2003) estimated the sample size that is required for dependent agreement studies by adapting the GEE2 approach for modeling dependent kappa statistics. Note that Klar et al. (2000); Williamson et al. (2000); Gonin et al. (2000); Lipsitz et al. (2001, 2003) stressed the fact that it is important to overfit the marginal models to avoid bias in the estimation of the agreement index.

The modeling techniques presented in this chapter were limited to the case of two or more raters. They have not yet been generalized to the case of two groups of raters or an isolated rater and a group of raters. Further research is needed on this topic.

Conclusion

This work has focused on the agreement between raters in various situations. A short review of agreement indexes for quantitative scales was provided in the first chapter, with particular emphasis on indexes also applicable to qualitative scales. Actually, Lin et al. (2007) proposed a unified approach for categorical and continuous data. The main corpus of this work, however, was devoted to the measurement of agreement between two or more raters on a categorical scale. Preference was given to agreement coefficients belonging to the kappa-like family. As clearly explained by Kraemer (1992), agreement on a categorical scale is often assessed by (1) defining some measure of pairwise agreement, (2) giving to each item an agreement score equal to the pairwise agreement measure averaged over all pairs of raters/ratings, (3) averaging the agreement scores over items, and (4) assessing how the agreement scores relate to what one would define as random and ideal agreement in that dataset.

One crucial point is to provide a clear definition of 'perfect agreement'. While it is unambiguous for two raters (they agree or do not agree), this is not the case when agreement is searched between several raters (where agreement can be defined on a continuum beginning with agreement between a pair of raters to agreement between all raters) or between groups of raters. In the latter case, the issue arises as to whether agreement within groups is needed to have agreement between groups. Another question is what is meant by chance agreement. For example, for Cohen's kappa coefficient (Cohen, 1960), chance agreement was defined as agreement between the two raters under the independence assumption. By contrast, Kraemer (1979) used the additional assumption of homogeneous marginal distribution for the intraclass kappa coefficient and Bennett et al. (1954) the additional assumption

of uniform marginal distribution for both raters. Different definitions of 'perfect agreement' and 'chance agreement' lead to different indexes and possibly to different conclusions.

Once clear definitions of 'perfect agreement' and 'chance agreement' have been specified, a raw measure of agreement (p_o) and the corresponding measure obtained by pure chance (p_e) are derived. Finally, an agreement coefficient is constructed, $\hat{\kappa} = (p_o - p_e)/(p_m - p_e)$, with the property of having a value equal to 1 when agreement is perfect. In this formula, p_m corresponds to the value of p_o under the definition of perfect agreement adopted. Based on this principle, Cohen's kappa coefficient (Cohen, 1960), the weighted kappa coefficient (Cohen, 1968) and the intraclass kappa coefficient (Kraemer, 1979) were reviewed. Then, the intraclass kappa coefficients based on one-way (Fleiss, 1971) and two-way (Davies and Fleiss, 1982) ANOVA models and the g-wise agreement indexes (Conger, 1980) were introduced to quantify the agreement between several raters. Finally, Schouten (1982) and Vanbelle and Albert (2009a,b) proposed agreement indexes for quantifying the agreement between an isolated rater and a group of raters or between two groups of raters. The question of "what is meant by agreement between two groups of raters?" remains a subject of debate. Vanbelle and Albert (2009a,b) offered an alternative proposal to existing methods, such as the consensus method and Schouten's approach (Schouten, 1982). The proposed agreement coefficients were defined on less restrictive definitions of perfect agreement than in Schouten's approach, in the sense that agreement within each group of raters was not requested to reach agreement between two groups of raters. The consensus method, consisting in summarizing the responses of each group of raters in a unique quantity, should be avoided since the information about the dispersion of the responses within each group of raters is erased. Moreover, the agreement index resulting from the consensus method is generally overestimated, because items without consensus are merely discarded from the statistical analysis. We are aware that the kappa-like family is not the unique issue in the study of agreement between raters but this remains the family of coefficients mostly used in practice at the moment. Alternative approaches suggest the use of log-linear models, Rash models or mixture models (Schuster, 2004) but also of other coefficients like the tetrachoric correlation coefficient (Pearson, 1900).

After having defined an agreement index belonging to the kappa-like family, practical interpretation of the values taken by the coefficient should be provided. For values like 0 (agreement due to chance) or 1 (perfect agreement), there is not much discussion, while for values between 0 and 1 the interpretation remains open. Landis and Koch (1977b) qualified the strength of agreement (from "poor" to "almost perfect") according to values taken by Cohen's kappa coefficient. Although widely

used in practice, this classification should be avoided because the qualification is arbitrary. A better approach would consist in the determination of a lower bound for the agreement index. This can be achieved by calculating the standard error of the estimated agreement index and by determining a 95% confidence interval for the unknown coefficient. While several methods have been developed over the years in this respect for agreement indexes between two raters, it is not the case for the other coefficients of agreement. This prompted us to propose the Jackknife technique to estimate sampling variability. Another aspect of interpretation that requires clarification is that concerning the use of weights. Weighted kappa coefficients are widely used for ordinal scales. Weights can be arbitrarily defined but traditionally linear or quadratic forms are applied. We gave an interpretation for the linear weighting scheme (Vanbelle and Albert, 2008) while (Schuster, 2004) proposed an interpretation for quadratic weights. However, guidelines for choosing one or the other type of weighting scheme are still missing. Further research is therefore needed on this important topic.

While the first part of the present work was devoted to quantification of agreement in various contexts, the second part was dedicated to hypothesis testing and modeling. Asymptotic and exact statistical tests for agreement assessment were provided. Exact methods are being preferred for small sample sizes because the distribution of kappa coefficients is not symmetric. When testing for differences between several agreement coefficients, we made a clear distinction between the paired and unpaired cases. Independent agreement indexes are obtained when considering independent samples of items, while dependent agreement indexes are obtained when considering the same sample of items but possibly different raters. We spotted only one asymptotic method in the literature (Fleiss, 1981) for comparing several independent agreement indexes. A search for exact methods may be another option for future research. The bootstrap method developed by McKenzie et al. (1996) to compare two dependent agreement indexes was generalized to the case of several raters by Vanbelle and Albert (2008). The method is simple to apply but suffers from the known drawbacks of the bootstrapping, namely that different results might be obtained when the entire bootstrap distribution is not determined.

Finally, methods to model agreement indexes according to categorical and continuous covariates between two raters were exposed. These methods are based on or are equivalent to the generalized estimating equations. A first set of equations is used to determine the marginal distribution of the responses of the raters and a second set to model the agreement coefficient according to covariates. The generalized estimating equations offer the advantage of adjusting for categorical and continuous covariates but the efficiency of parameter estimation highly depends on the correct specification of the model. These methods need to be expanded

to model agreement indexes between an isolated rater and a group of raters or between two groups of raters.

Some aspects of rater agreement were not discussed in this work like agreement coefficient developed for paired data (Oden, 1991; Schouten, 1993; Shoukri et al., 1995) and stratified agreement coefficients (Barlow et al., 1991). Our contribution to the vast domain of rater agreement has raised more questions than solving ones but it has hopefully opened new pathways for future research.

This work was based on the following papers:

- Vanbelle S., Massart V., Giet D. and Albert A. (2007) *Test de concordance de script: un nouveau mode d'établissement des scores limitant l'effet du hasard*, Pédagogie Médicale, 8, pp 71-81.
- Vanbelle S. and Albert A. (2008) *A bootstrap method for comparing correlated kappa coefficients*, Journal of Statistical Computation and Simulation, 78, pp 1009-1015.
- Vanbelle S. and Albert A. (2009a) *A note on the linearly weighted kappa coefficient for ordinal scales*, Statistical Methodology, 6, pp 157-163.
- Vanbelle S. and Albert A. (2009b) *Agreement between an isolated rater and a group of raters*, Statistica Neerlandica, 1, pp 82-100.
- Vanbelle S. and Albert A. (2009c) *Agreement between two groups of raters*, Psychometrika, in press.
- Collard A., Gelaes S., Vanbelle S., Bredart S., Defraigne J-O., Boniver J. and Bourguignon J-P. (2009) *Reasoning versus knowledge retention and ascertainment throughout a PBL curriculum*, Medical education, to appear.

APPENDIX A

Data sets

A.1 Chapter 1

The serum gentamicin concentrations ($\mu mol/L$) measured twice with the EMIT and the FIA methods on 56 specimens are given in Table A.1.

Table A.1. Serum gentamicin concentrations ($\mu mol/L$) measured with the EMIT and the FIA methods on 56 specimens

Specimen	EMIT		FIA	
	Measure 1	Measure 2	Measure 1	Measure 2
1	2.5	2.6	2.4	2.3
2	4.4	3.9	4.2	3.9
3	2.4	2.7	2.1	1.9
4	4	3.8	3.6	3.5
5	4.9	4.1	3.1	2.8
6	2.4	2.2	2.3	2.4
7	2.9	3.3	3.4	3.3
8	8.5	6.7	12	8.5
9	2.7	2.6	2.9	3
10	2.6	3.1	2.8	2.7
11	2.4	3.3	2.8	2.6
12	3.2	2.6	3.4	3.4
13	5.3	3.7	6.8	6.7
14	2	2.4	2.3	3.5
15	4.7	4.7	4.1	4.1
16	2.7	2.9	2	2
17	6	4.2	4.5	4.6
18	3.8	3.6	3.6	3.7
19	3.1	3.3	4.8	4.6
20	7.9	5.6	4.2	3.6
21	1.4	3.2	2.4	2.3
22	2.1	1.8	2.3	2.3
23	2.6	1	2.1	2.1
24	3	5.2	1.4	2.7
25	5	4.5	4.8	4.9
26	2	2.2	2.8	2.4
27	1	1	1.2	1.2
28	11	11.2	10.6	11.3
29	1.2	1.3	1.8	0.8
30	2.3	1	3.9	4.2
31	6.4	5.9	5.3	5.8
32	4.8	4	5.8	6.3
33	2.2	1.8	2.1	2.2
34	3.7	3.2	4.2	4.1
35	7.4	6.7	9.2	9.4
36	1	2.5	2.4	2.6
37	6.8	8.5	7.2	7.4
38	1	1.4	1.8	1.9
39	1	1	2.5	2.5
40	5	4.5	6.4	6.7
41	2.1	1.1	3.2	3.4
42	5.4	6.2	5.8	6.1
43	10.4	8.6	9.6	10.1
44	6.8	6.9	7.6	7.9
45	7.3	8	5.2	6.4
46	7.6	6.1	6.2	6.2
47	10.5	11.5	10.2	10.2
48	9.8	11.5	10.5	10.5
49	14.5	13.5	12.8	12.4
50	16.5	12.5	13.2	13.2
51	19	16.5	15.5	15.8
52	19	17.5	15.7	16.2
53	12.8	11.9	12.5	12.9
54	17.4	13.3	15.7	16
55	11	10.8	12.3	11.7
56	13.9	14.2	13.5	13.8

A.2 Chapter 3

The data of Williams (1976) relative to the classification of 28 specimens for syphilis serology on a 3-category scale (NR=Non reactive, BL=Bordeline, RE=Reactive) by 2 individual laboratories (L and H) and 3 reference laboratories are presented in Table A.2.

Table A.2. Classification of 28 specimens for syphilis serology on a 3-category scale (NR=Non reactive, BL=Bordeline, RE=Reactive) by 2 individual laboratories (L and H) and 3 reference laboratories (Data from Williams (1976))

	Participant		Reference		
Specimen	L	H^*	R1	R2	R3
1	RE	RE	RE	RE	RE
2	RE	RE	RE	RE	RE
3	BL	NR	NR	NR	NR
4	BL	NR	NR	NR	NR
5	BL	NR	NR	NR	NR
6	RE	RE	RE	RE	RE
7	BL	NR	NR	NR	NR
8	RE	RE	RE	RE	RE
9	NR	NR	NR	NR	NR
10	NR	NR	NR	NR	NR
11	RE	RE	RE	RE	RE
12	RE	BL	RE	BL	BL
13	RE	RE	RE	RE	RE
14	RE	BL	RE	BL	BL
15	RE	RE	RE	RE	RE
16	RE	BL	RE	NR	BL
17	RE	BL	RE	NR	BL
18	RE	RE	RE	RE	RE
19	RE	RE	RE	RE	RE
20	BL	NR	BL	NR	NR
21	RE	RE	RE	RE	RE
22	BL	NR	NR	NR	NR
23	BL	NR	BL	NR	NR
24	BL	NR	BL	NR	NR
25	RE	RE	RE	RE	RE
26	NR	NR	NR	NR	NR
27	RE	RE	RE	RE	RE
28	NR	NR	NR	NR	NR

*Hypothetical participant (see text)

APPENDIX B

Asymptotic and exact methods

B.1 Introduction

First, the multivariate Delta method will be exposed in the general case and in the particular case of multinomial distribution in order to determine the standard error of Cohen's kappa, intraclass kappa and weighted kappa coefficients. Then, sampling methods such as Jackknife, bootstrap and Monte Carlo approximation will be introduced.

B.2 Multivariate Delta method

B.2.1 General case

Let $\boldsymbol{\theta}$ be a vector of population parameters of dimension $T \times 1$: $\boldsymbol{\theta}=(\theta_1, \ldots, \theta_T)'$; $\hat{\boldsymbol{\theta}}_n$ be a vector of estimators of dimension $T \times 1$ of the vector of $\boldsymbol{\theta}$ for a size N: $\hat{\boldsymbol{\theta}}_N = (\hat{\theta}_{N1}, \ldots, \hat{\theta}_{NT})'$.

Suppose that $\hat{\boldsymbol{\theta}}_N$ is asymptotically normally distributed, i.e.,

$$\mathcal{L}[\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta})] \longrightarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{B.1}$$

where $\mathcal{L}$ represents convergence in law and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the $T \times T$ asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_N$. This matrix is singular if $\hat{\boldsymbol{\theta}}_N$ has a distribution included in a sub-space of the T-dimensional space.

Suppose that $\boldsymbol{f}$ is a function defined on an open subspace of the T-dimensional space and taking values in a R-dimensional space,

$$\boldsymbol{f} : \mathbb{R}^T \ \rightarrow \ \mathbb{R}^R \ : \ \boldsymbol{\theta} \mapsto \boldsymbol{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_R(\boldsymbol{\theta}))'.$$

Assume that $\boldsymbol{f}$ is at least one time differentiable in $\boldsymbol{\theta}$, i.e.

$$f_i(\boldsymbol{x}) = f_i(\boldsymbol{\theta}) + \sum_{j=1}^{T} (x_j - \theta_j) \frac{\partial f_i}{\partial x_j} \Bigg| \ \boldsymbol{x} = \boldsymbol{\theta} \ + o(||\boldsymbol{x} - \boldsymbol{\theta}|\,|) \text{ if } \boldsymbol{x} \rightarrow \boldsymbol{\theta} \text{ for } i = 1, \ldots, R.$$

If $\left(\dfrac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right)$ is the $R \times T$ matrix whose element (i, j) is the partial derivative of f_i with respect to the jth element of $\boldsymbol{x} = (x_1, \ldots, x_T)'$ evaluated in $\boldsymbol{x} = \boldsymbol{\theta}$, i.e.,

$$\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right)_{ij} = \frac{\partial f_i}{\partial x_j} \Bigg| \ \boldsymbol{x} = \boldsymbol{\theta},$$

Then,

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{\theta}) + \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right) (\boldsymbol{x} - \boldsymbol{\theta}) + o(||\boldsymbol{x} - \boldsymbol{\theta}|\,|) \text{ if } \boldsymbol{x} \rightarrow \boldsymbol{\theta}. \tag{B.2}$$

Bishop et al. (1975) have shown that

Theorem B.2.1. *If $\hat{\boldsymbol{\theta}}_N$, $\boldsymbol{\theta}$ and $\boldsymbol{f}$ are defined as above and (B.1) and (B.2) hold, then, the asymptotic distribution of $\boldsymbol{f}(\hat{\boldsymbol{\theta}}_N)$ is given by*

$$\mathcal{L}[\sqrt{N}(\boldsymbol{f}(\hat{\boldsymbol{\theta}}_N) - \boldsymbol{f}(\boldsymbol{\theta}))] \ \longrightarrow \ \mathcal{N}\left(\mathbf{0}, \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right) \Sigma(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \right)' \right). \tag{B.3}$$

B.2.2 Particular case: multinomial distribution

Simplifications occur in the multinomial case. Suppose that we have a $K \times K$ contingency table and that the cell counts $(n_{11}, \ldots, n_{1K}, \ldots, n_{K1}, \ldots, n_{KK})'$ follow a multinomial distribution with probability cells

$$\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1K}, \ldots, \pi_{K1}, \ldots, \pi_{KK})'.$$

We have

$$\sum_{i=1}^{K} \sum_{j=1}^{K} n_{ij} = N.$$

Let $\boldsymbol{p} = (p_{11}, \ldots, p_{1K}, \ldots, p_{K1}, \ldots, p_{KK})'$ be the vector of sample proportions where $p_{ij} = n_{ij}/K$.

Let the ith observation be $\boldsymbol{y_i} = (y_{i11}, \ldots, y_{i1K}, \ldots, y_{iK1}, \ldots, y_{iKK})$ where $y_{ijl} = 1$ if item i is placed in category j by rater 1 and in category l by rater 2, $y_{ijl} = 0$ otherwise. We have

$$p_{jl} = \frac{1}{N}\sum_{i=1}^{N} y_{ijl},\ \sum_{j=1}^{K}\sum_{l=1}^{K} y_{ijl} = 1 \text{ and } y_{ijl}y_{imr} = 0 \text{ if } j \neq m \text{ or if } l \neq r.$$

Moreover, $E(y_{ijl}) = \pi_{jl} = E(y_{ijl}^2)$ and $E(y_{ijl}y_{imr}) = 0$ if j$\neq$ m or if $l \neq r$. So, $E(\boldsymbol{y_i}) = \boldsymbol{\pi}$ and $cov(\boldsymbol{y_i}) = \boldsymbol{\Sigma}$ $(i = 1, \ldots, N)$ where $\boldsymbol{\Sigma} = (\sigma_{jl})$ with $\sigma_{jj} = var(y_{ijl}) = \pi_{jl}(1 - \pi_{jl})$ and $\sigma_{jl} = cov(y_{ijl}, y_{imr}) = -\pi_{jl}\pi_{mr}$ if $j \neq m$ or if $l \neq r$.

If $diag(\boldsymbol{\pi})$ denotes the diagonal matrix with the elements of $\boldsymbol{\pi}$, the matrix $\boldsymbol{\Sigma}$ is determined by

$$\boldsymbol{\Sigma} = diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi'}. \tag{B.4}$$

Thus, we have

$$cov(\boldsymbol{p}) = \frac{1}{N}(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi'}).$$

This matrix is singular since $\sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij} = 1$.
The multivariate central-limit theorem (Rao, 1973) implies that

$$\mathcal{L}\left[\sqrt{N}\left[\boldsymbol{p} - \boldsymbol{\pi}\right]\right] \longrightarrow \mathcal{N}\left[\boldsymbol{0}, diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi'}\right].$$

If $g(t_{11}, \cdots, t_{1K}, \cdots, t_{K1}, \cdots, t_{KK})$ is a differentiable function and

$$\phi_{ij} = \frac{\partial g}{\partial \pi_{ij}} \quad i, j = 1, \ldots, K$$

is $\dfrac{\partial g}{\partial \boldsymbol{t}}$ evaluated in $\boldsymbol{t} = \boldsymbol{\pi}$, the Delta method implies that

$$\mathcal{L}\left[\sqrt{N}\left[g(\boldsymbol{p}) - g(\boldsymbol{\pi})\right]\right] \longrightarrow \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\phi}'\left[(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi'})\right]\boldsymbol{\phi}\right)$$

where $\phi' = (\phi_{11}, \cdots, \phi_{1K}, \cdots, \phi_{K1}, \cdots, \phi_{KK})$.

The asymptotic covariance matrix of $g(\boldsymbol{p})$ is thus equal to

$$\boldsymbol{\phi'}diag(\boldsymbol{\pi})\boldsymbol{\phi} - (\boldsymbol{\phi'}\boldsymbol{\pi})^2 = \sum_{i=1}^{K}\sum_{j=1}^{K}\pi_{ij}\phi_{ij}^2 - (\sum_{i=1}^{K}\sum_{j=1}^{K}\pi_{ij}\phi_{ij})^2. \tag{B.5}$$

To determine the sampling variance of a kappa coefficient, the function $g(\boldsymbol{p}) = \kappa(\boldsymbol{p})$ is considered.

B.3 Jackknife method

The Jackknife is a statistical method introduced by Quenouille (1956) and further developed by Tukey (1958) for reducing the bias in an estimator of some population parameter and for obtaining an estimate of the standard error of the improved estimator.

Suppose one has classified N items into K mutually exclusive and exhaustive categories, and that the number of items assigned to the respective categories are $n_1, \cdots, n_K$, with $\sum_{i=1}^{K} n_i = N$. Consider some function $F = F(n_1, \cdots, n_K)$ that is to serve as an estimator of a parameter θ, but suppose that F is such that

$$E(F) = \theta + O(\frac{1}{N})$$

The Jackknife estimator, $\widetilde{F}$, will be such that

$$E(\widetilde{F}) = \theta + \frac{1}{N^2}.$$

Let F_{-i} be the value of the function when one unit is deleted from the ith category and define the K so-called *pseudovalues* $\widetilde{F}_i = NF - (N-1)F_{-i}$. The Jackknife estimator of θ, $\widetilde{F}$ is the weighted average of the pseudovalues,

$$\widetilde{F} = \frac{1}{N} \sum_{i=1}^{K} n_i \widetilde{F}_i. \tag{B.6}$$

and the estimate variance of $\widetilde{F}$ is

$$var(\widetilde{F}) = \frac{1}{N(N-1)} \sum_{i=1}^{K} n_i (\widetilde{F}_i - \widetilde{F})^2. \tag{B.7}$$

If N is 'large', inferences about θ may be based on the fact that $\widetilde{F}$ is approximately normally distributed about θ with a standard error of $\sqrt{(var(\widetilde{F}))}$.

The Jackknife is useful when no explicit formula is available for the variance of F or when the variance formula is complicated.

B.4 Bootstrap and Monte Carlo approximation

B.4.1 Bootstrap

Assume a data set has N observations. The bootstrap first forms a finite population by giving each observation a probability $1/N$. From this finite population,

there are N^N possible samples of size N obtained with replacement. For each of these samples, value of the statistic under study can be computed, given N^N such values. The empirical distribution of the statistic is the *exact bootstrap* distribution. The end points of the 95% bootstrap confidence interval are obtained by the *percentile method*, i.e., by taking the 2.5th and the 97.5th percentiles from the bootstrap distribution.

B.4.2 Monte Carlo approximation

Unfortunately, for most problems, calculating the *exact bootstrap distribution* is computationally prohibitive. Therefore, a Monte Carlo approximation is used. The Monte Carlo bootstrap draws B samples of size N with replacement from the original data set. For each of these B samples, the value of the statistic is calculated, giving B possibly different values. The end points of the 95% Monte Carlo bootstrap confidence interval are obtained by taking the 2.5th and the 97.5th percentiles from the empirical distribution of the B values of κ.

APPENDIX C

Generalized linear models

C.1 Introduction

In this Appendix, the generalized exponential family is outlined. The first two moments of such random variable are determined. Generalized linear models and logistic regressions are introduced for binary and ordinal variables (Nelder and Wedderburn, 1972). It is a generalization of the linear model defined for normally distributed populations to the generalized exponential family. Maximum likelihood equations for such model and iterative Fisher scoring method are exposed.

C.2 Generalized exponential family

C.2.1 Definition

Let Y be a random variable and suppose that $(y_1, \cdots, y_N)$ represent the values of N independent observations of the random variable Y.

Y is a generalized exponential random variable if the density of each y_i $(i = 1, \cdots, N)$ with respect to a Lebegue or count measure λ can be written as

$$f(y_i; \theta_i, \phi) = exp\left[(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\right]. \tag{C.1}$$

The parameter θ_i is called the *canonical parameter* and the parameter ϕ the *dispersion parameter*. The functions $a(\phi)$ and $b(\theta_i)$ are supposed to be at least two times continuously differentiable in the parameter space.

C.2.2 Two first moments of Y

Consider the contribution of the ith observation to the likelihood logarithm

$$l(\theta_i, \phi; y_i) = \ln(f(y_i; \theta_i, \phi)) = [(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)]. \tag{C.2}$$

We have

$$\partial l/\partial\theta_i = [y_i - b'(\theta_i)]\,/a(\phi) \tag{C.3}$$

$$\partial^2 l/\partial\theta_i^2 = -b''(\theta_i)/a(\phi) \tag{C.4}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are, respectively, the two first derivatives of b evaluated at θ_i.

It is easily shown that if the conditions of permutation of the integration and derivation operators are satisfied,

$$E\left(\frac{\partial l}{\partial\theta_i}\right) = 0 \tag{C.5}$$

and

$$-E\left(\frac{\partial^2 l}{\partial\theta_i^2}\right) = E\left(\frac{\partial l}{\partial\theta_i}\right)^2. \tag{C.6}$$

It implies, with respect to (C.3) and (C.5), if μ_i represents $E(Y_i)$,

$$\mu_i = b'(\theta_i) \tag{C.7}$$

and with respect to (C.4) and (C.6),

$$E\left[(y_i - b'(\theta_i))^2\,/a^2(\phi)\right] = var(Y_i)/a^2(\phi) = b''(\theta_i)/a(\phi).$$

Finally,

$$var(Y_i) = b''(\theta_i)a(\phi). \tag{C.8}$$

C.3 Systematic component and link function

Let $x_{i1}, \cdots, x_{ip}$ be the values of p covariates relative to the ith observation. The *systematic component* links the unknown parameter to the covariates using a linear predictor

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij},\ i = 1, \cdots, N$$

or under matricial notation,

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} \tag{C.9}$$

where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_N)'$ is a vector of linear predictors, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is a vector of unknown parameters and $\boldsymbol{X}$ is a $N \times p$ matrix of covariates.

We also defined

$$\eta_i = g(\mu_i) \tag{C.10}$$

where g is a strictly monotone function at least two times continuously differentiable with respect to $\mu_1, \cdots, \mu_N$. g is called the *link function* and the function g for which $g(\mu_i) = \theta_i$ is called the *canonical link*.

C.4 Estimation of the parameters

For N independent observations, the likelihood logarithm is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \ln(f(y_i; \theta_i, \phi)) = \sum_{i=1}^{N} l_i \tag{C.11}$$

where $l_i = l(\theta_i, \phi, y_i),\ (i = 1, \cdots, N)$.

The maximum likelihood equations are obtained by calculating

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{C.12}$$

We have $\dfrac{\partial l_i}{\partial \theta_i} = [y_i - b'(\theta_i)]/a(\phi)$, $\mu_i = b'(\theta_i)$ and $var(Y_i) = b''(\theta_i)a(\phi)$. Therefore,

$$\frac{\partial l_i}{\partial \theta_i} = (y_i - \mu_i)/a(\phi) \text{ and } \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{var(Y_i)}{a(\phi)}.$$

Moreover,

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} \text{ thus } \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

This implies

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)}\frac{a(\phi)}{var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij}. \tag{C.13}$$

After some elementary algebraic manipulations, we obtain the maximum likelihood equations

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)}{var(Y_i)} x_{ij}\frac{\partial \mu_i}{\partial \eta_i} = 0,\ j = 1, \cdots, p \tag{C.14}$$

or, under matricial notation,

$$\boldsymbol{X}'\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0} \tag{C.15}$$

where $\boldsymbol{\Delta}$ is the diagonal matrix composed by the elements $(\dfrac{\partial \mu_i}{\partial \eta_i}\dfrac{1}{var(Y_i)})$.

Such equations do not have analytical solution. The iterative method used to fit the generalized linear model is the Fisher iterative method described in next Section.

Convergence rate of $\hat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}$ depends of the information matrix.

We have

$$\begin{aligned} -E\left(\frac{\partial^2 l_i}{\partial\beta_h\partial\beta_j}\right) &= E\left[\left(\frac{\partial l_i}{\partial\beta_h}\right)\left(\frac{\partial l_i}{\partial\beta_j}\right)\right] \\ &= E\left[\frac{(Y_i-\mu_i)x_{ih}}{var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\frac{(Y_i-\mu_i)x_{ij}}{var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right] \\ &= \frac{x_{ih}x_{ij}}{var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \end{aligned} \tag{C.16}$$

Thus,

$$-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_h\partial\beta_j}\right) = \sum_{i=1}^{N}\frac{x_{ih}x_{ij}}{var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{C.17}$$

Under matricial notation,

$$\boldsymbol{H} = \boldsymbol{X'WX} \tag{C.18}$$

where
$\boldsymbol{H} = E\left(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_h\partial\beta_j}\right)$ is called *information matrix*;
$\boldsymbol{W}$ is the diagonal matrix with $w_i = \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 / var(Y_i)$ on the main diagonal.

C.5 Fisher scoring method

Let $\boldsymbol{\beta}^{(m)}$ be the *mth* approximation of the estimation $\hat{\boldsymbol{\beta}}$ obtained by the Newton-Raphson method where the matrix of the second derivatives is replaced by its expectation $-\boldsymbol{H}$.

We have

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\boldsymbol{H}^{(m)})^{-1}\boldsymbol{q}^{(m)} \tag{C.19}$$

where
$\boldsymbol{H}$ is the non-singular matrix defined in (C.18) ;
$\boldsymbol{q}$ is the vector with elements $\frac{\partial L(\boldsymbol{\beta})}{\partial\beta_j}$;
$\boldsymbol{H}^{(m)}$ and $\boldsymbol{q}^{(m)}$ are $\boldsymbol{H}$ and $\boldsymbol{q}$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$.

Regarding Equations C.14 and C.17,

$$\boldsymbol{H}^{(m)}\boldsymbol{\beta}^{(m)} + \boldsymbol{q}^{(m)} = \sum_{j=1}^{p}\left(\sum_{i=1}^{N}\frac{x_{ih}x_{ij}}{var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\beta_j^{(m)}\right) + \sum_{i=1}^{N}\frac{(y_i-\mu_i^{(m)})}{var(Y_i)}x_{ik}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)$$

where μ_i and $\left(\frac{\partial\mu_i}{\partial\eta_i}\right)$ are evaluated at $\boldsymbol{\beta}^{(m)}$.

So,

$$\boldsymbol{H}^{(m)}\boldsymbol{\beta}^{(m)} + \boldsymbol{q}^{(m)} = \boldsymbol{X'W}^{(m)}\boldsymbol{Z}^{(m)} \tag{C.20}$$

where $\boldsymbol{W}^{(m)}$ is $\boldsymbol{W}$ evaluated at $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{Z}^{(m)}$ is composed by the elements

$$\begin{aligned} z_i^{(m)} &= \sum_{j=1}^{p}x_{ij}\beta_j^{(m)} + (y_i-\mu_i^{(m)})\left(\frac{\partial\eta_i^{(m)}}{\partial\mu_i^{(m)}}\right) \\ &= \eta_i^{(m)} + (y_i-\mu_i^{(m)})\left(\frac{\partial\eta_i^{(m)}}{\partial\mu_i^{(m)}}\right). \end{aligned} \tag{C.21}$$

Finally, Fisher's equations (C.19) have following form

$$\begin{aligned} \boldsymbol{H}^{(m)}\boldsymbol{\beta}^{(m+1)} &= \boldsymbol{H}^{(m)}\boldsymbol{\beta}^{(m)} + \boldsymbol{q}^{(m)} \\ \boldsymbol{X'W}^{(m)}\boldsymbol{X}\boldsymbol{\beta}^{(m+1)} &= \boldsymbol{X'W}^{(m)}\boldsymbol{Z}^{(m)}. \end{aligned} \tag{C.22}$$

If the solution exists and is uniquely defined, the solution of the equations is

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X'W}^{(m)}\boldsymbol{X})^{-1}\boldsymbol{X'W}^{(m)}\boldsymbol{Z}^{(m)}. \tag{C.23}$$

The vector $\boldsymbol{Z}^{(m)}$ represents, in that expression, a linearized form of the link function in $\boldsymbol{\mu}$ evaluated at $\boldsymbol{y}$.

$$\begin{aligned} g(y_i) &\sim g(\mu_i) + (y_i-\mu_i)g'(\mu_i) \\ &\sim \eta_i + (y_i-\mu_i)\frac{\partial\eta_i}{\partial\mu_i} = z_i. \end{aligned} \tag{C.24}$$

It is important to remark that Fisher scoring method does not always converge. In practice, we consider $\epsilon > 0$ and a maximal number of iterations M. The iterative procedure stops if

$$\forall j \in \{1,\cdots,p\} \quad \frac{\left\|\beta_j^{(m+1)} - \beta_j^{(m)}\right\|}{\|\beta^{(m)}\|} < \epsilon$$

or if $m = M$.

C.6 Logistic regression

C.6.1 Binary logistic regression

Many categorical variables only possess two categories. Each observation on each item may be a success (value 1) or a failure (value 0). For binary random variables, the Bernoulli distribution specifies the probabilities $P(Y = 1) = \pi$, $P(Y = 0) = 1 - \pi$. It results that $E(Y) = \pi$ and $var(Y) = \pi(1 - \pi)$. When Y_i has a Bernoulli distribution with parameter π_i, the density distribution is

$$f(y_i, \pi_i) = \exp\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + ln(1-\pi_i)\right] \tag{C.25}$$

for $y_i = 0$ and $y_i = 1$. This distribution takes place in the generalized exponential distribution. The canonical parameter, θ_i, is $\ln(\frac{\pi_i}{1-\pi_i})$. This term is called $logit(\pi_i)$.

C.6.1.1 Linear regression

For a binary response and one covariate, the linear regression model is

$$E(Y) = \pi(x) = \beta_0 + \beta_1 x. \tag{C.26}$$

When the observations y are independent, this model corresponds to the generalized linear model with identity link.

The major default of the model is the following. While proportions $\pi(x)$ have to be between 0 and 1, the linear function permits values on all the real line, i.e., values of π smaller than 0 or greater than 1. It is then proposed to take a non-linear relation between $\pi(x)$ and x. The appropriate model is introduced in the next Section.

C.6.1.2 Logistic regression for a single covariate

For one covariate, the proposed model is the following:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{C.27}$$

called the *logistic regression* function.

When $x \to \infty$, $\pi(x) \downarrow 0$ if $\beta_1 < 0$;
$\pi(x) \uparrow 1$ if $\beta_1 > 0$.

The model is represented by a sigmoidal curve and possesses the properties of a continuous repartition function. The link function for which the logistic regression model is a generalized linear model is the logit link

$$\ln(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_1 x. \tag{C.28}$$

Suppose they are g groups of n_i observations ($i = 1, \cdots, g$). Let y_i be the ith observation of the study binomial random variable.

$$\ln\left(\frac{y_i}{n_i - y_i}\right) \tag{C.29}$$

is not defined when $y_i = 0$ or $y_i = n_i$. Therefore, the *empirical logit*, which is a biased estimator of the real logit, is sometimes used instead:

$$\ln\left(\frac{y_i + 1/2}{n_i - y_i + 1/2}\right). \tag{C.30}$$

The generalization of the logit function to several covariates is simple. Let $\boldsymbol{x} = (x_1, \cdots, x_p)'$ be the values of p covariates. The logistic regression model is:

$$\ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{C.31}$$

C.6.1.3 Likelihood estimators

Let $(y_1, \cdots, y_N)$ be the values of N binary variables. We suppose that those random variables are independent and possess a Bernoulli distribution.
Let $\boldsymbol{x}_i = (x_{i0}, \cdots, x_{ip})$ be the ith set of p covariates, $i = 1, \cdots, I$ and $x_{i0} = 1$. When covariates are continuous, a different set of covariates may exist for each subject and $I = N$.

The logistic regression model is:

$$\pi(\boldsymbol{x}_i) = \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}. \tag{C.32}$$

Let n_i be the number of observations for a fixed value of $\boldsymbol{x}_i = (x_{i0}, \cdots, x_{ip})$. Y_i is a random variable counting number of successes. The random variables Y_i, $(i = 1 \cdots, I)$ are independent binomial random variables where $E(Y_i) = n_i \pi(\boldsymbol{x_i})$ and $n_1 + \cdots + n_I = N$.

The probability density is

$$f(y_i, \pi) = \exp\left[y_i \ln\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right) + n_i \ln(1 - \pi(\boldsymbol{x}_i))\right]. \tag{C.33}$$

Thus, we have

$$\eta_i = \theta_i; \quad \mu_i = n_i\pi(\boldsymbol{x}_i); \quad b(\theta_i) = -n_i \ln(1-\pi(\boldsymbol{x}_i));$$

$$var(Y_i) = n_i\pi(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i)) \text{ et } a(\phi) = 1.$$

$$\frac{\partial\theta_i}{\partial\mu_i} = \frac{1}{n_i\pi(\boldsymbol{x}_i)} + \frac{1}{n_i(1-\pi(\boldsymbol{x}_i))} = \frac{1}{n_i\pi(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i))}.$$

So, the maximum likelihood equations are

$$\sum_{i=1}^{I}(y_i - n_i\pi_i)x_{ij} = 0, \quad j = 0, \cdots, p \tag{C.34}$$

or, under matricial form, if $\boldsymbol{X}$ is the $I\times(p+1)$ matrix composed by $\{x_{ij}\}$, Equation C.34 has the form

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{\mu} \tag{C.35}$$

Information matrix can then be written as

$$H_{kj} = \sum_{i=1}^{I} x_{ik}x_{ij}var(Y_i) \tag{C.36}$$

or

$$\boldsymbol{H} = \boldsymbol{X}'\boldsymbol{diag}(var(Y_i))\boldsymbol{X}. \tag{C.37}$$

For the logistic regression, the maximum likelihood estimators exist and are uniquely defined except under limit cases (see Wedderburn (1976), Albert and Anderson (1984) and Lesaffre and Albert (1989) for more detail). The maximum likelihood equations are non linear functions of the maximum likelihood estimations $\hat{\boldsymbol{\beta}}$. Resolution of those equations can be done using the Newton-Raphson iterative method.

C.6.2 Ordinal logistic regression

Different models exist for ordinal data. Only the most popular in the modelisation of agreement data will be exposed here (see Hosmer and Lemeshow (2000) for the other models).

Let Y be an ordinal random variable, which may take $K+1$ values, noted $0, \cdots, K$ and $\boldsymbol{x} = (x_1, \cdots, x_p)'$ a vector of p covariates.

Let

$$P[Y = k|\boldsymbol{x}] = \phi_k(\boldsymbol{x}). \tag{C.38}$$

Suppose we may to compare the probabilities $P[Y \leq k|\boldsymbol{x}]$ and $P[Y > k|\boldsymbol{x}]$.

We define

$$\begin{aligned} c_k(\boldsymbol{x}) &= \ln\left[\frac{P[Y \leq k|\boldsymbol{x}]}{P[Y > k|\boldsymbol{x}]}\right] \\ &= \ln\left[\frac{\phi_0(\boldsymbol{x}) + \cdots + \phi_k(\boldsymbol{x})}{\phi_{k+1}(\boldsymbol{x}) + \cdots + \phi_K(\boldsymbol{x})}\right] \\ &= \ln\left[\frac{\gamma_k(\boldsymbol{x})}{1 - \gamma_k(\boldsymbol{x})}\right] \\ &= \tau_k - \boldsymbol{x}'\boldsymbol{\beta} \end{aligned} \tag{C.39}$$

for $k = 0, \cdots, K-1$, where $\gamma_k = \phi_0(\boldsymbol{x}) + \cdots + \phi_k(\boldsymbol{x})$ and τ_k is the intercept.

That model is called the *linear cumulative logistic model* and possess the following property

$$\ln\left(\frac{\gamma_k(\boldsymbol{x}_1)}{1 - \gamma_k(\boldsymbol{x}_1)}\right) - \ln\left(\frac{\gamma_k(\boldsymbol{x}_2)}{1 - \gamma_k(\boldsymbol{x}_2)}\right) = \boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2) \tag{C.40}$$

i.e., the difference between the two logistic is independent of the category k.

Mc Cullagh (1980) has defined the maximum likelihood equations and the iterative Newton-Raphson procedure for the ordinal logistic regression.

C.7 Goodness of fit

C.7.1 The goodness of fit statistic

A quality criterion of the fitting of a logistic (ordinal) regression is given by

$$G^2 = -2\sum_{i=1}^{N}[l_i(\hat{\mu}_i) - l_i(y_i)] \tag{C.41}$$

where l_i are defined by Equation C.11) This statistic G^2 is called *the likelihood ratio*. It can be shown that G^2 follows asymptotically a chi-square distribution on $N - p$ degrees of freedom where p is the number of estimated parameters.

Consider two hierarchical models

$M_0 :\ \ln(\frac{\pi_i}{1-\pi_i}) = \beta_0$ and $M_p :\ \ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \quad (i = 1, \cdots, N)$.

Let $G^2(M_0)$ and $G^2(M_p)$ be the likelihood ratios relative to the models M_0 and M_p respectively.

It can be shown that

$$G^2(M_p|M_0) = G^2(M_0) - G^2(M_p) \tag{C.42}$$

is asymptotically distributed as a chi-square statistic with p degrees of freedom. $G^2(M_p|M_0)$ appreciates the goodness of fit of the regression involving p covariates.

C.7.2 Standard error of the parameters

It is well known that the standard error of the estimator $\hat{\beta}_j$, $j = 0, \cdots, p$ is given by

$$SE(\hat{\beta}_j) = (H^{-1})_{jj}, \ (j = 1, \cdots, p) \tag{C.43}$$

where $\boldsymbol{H}^{-1}$ is the inverse of the information matrix and that

$$Z(\hat{\beta}_j) = \frac{\hat{\beta}_j - \beta_j}{SE(\beta_j)} \tag{C.44}$$

is approximately normally distributed with mean 0 and standard deviation 1.

It is then possible to test hypotheses

$$H_0 : \quad \beta_j = 0 \text{ vs } \quad H_1 : \quad \beta_j \neq 0 \quad (j = 0, \cdots, p)$$

at the α significance level by comparing $|Z(\hat{\beta}_j)|$ to $Q_Z(1 - \frac{\alpha}{2})$.

Several authors test H_0 with the Wald statistic defined by

$$\chi^2_{\beta_j} = Z^2(\beta_j) \tag{C.45}$$

H_0 is rejected at α significance level if $\chi^2_{\beta_j} > Q_{\chi^2}(1 - \alpha; 1)$ otherwise H_0 is not rejected.

C.8 Generalized estimating equations

Aim of this Section is to extend the generalized linear models to the case of paired data. Approach of Liang and Zeger (1986) will be developed. They defined generalized estimating equations (GEE) based on a *mean population* model.

Consider I blocs of paired data mutually independent.

Note

$$\boldsymbol{y} = (y_{11}, \cdots, y_{1n_1}, \cdots, y_{I1}, \cdots, y_{In_I}) = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_I)$$

the vector of observations,

$$\boldsymbol{\mu} = (\mu_{11}, \cdots, \mu_{1n_1}, \cdots, \mu_{I1}, \cdots, \mu_{In_I}) = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_I)$$

the mean vector and

$$\boldsymbol{X}_i = \begin{pmatrix} x_{i11} & \cdots & x_{i1p} \\ \vdots & & \vdots \\ x_{in_i1} & \cdots & x_{in_ip} \end{pmatrix}$$

the matrix $n_i \times p$ of covariates relative to the ith item ($i = 1, \cdots, I$) where $\boldsymbol{x}_{ij} = (x_{ij1}, \cdots, x_{ijp})$ is the vector of covariates corresponding to the jth observation of the ith item.

Suppose that the vector $\boldsymbol{y}$ is extracted from a generalized exponential population and that the following linear model exists:

$$\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}, \quad i = 1, \cdots, I$$

where $\boldsymbol{\eta}_i = (\eta_{i1}, \cdots, \eta_{in_i})'$ is a vector of linear predictors and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is a vector of unknown parameters. Moreover,

$$\eta_{im} = g(\mu_{im}) \quad i = 1, \cdots, I \quad m = 1, \cdots, n_i$$

is the link function.

Liang and Zeger (1986) approach needs hypotheses on the correlation nature between the paired data.

For example, a $n_i \times n_i$ correlation matrix proposed by Liang and Zeger (1986), noted $\boldsymbol{R}_i(\alpha)$, has the form

$$\boldsymbol{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha \\ \alpha & \cdots & \alpha & 1 \end{pmatrix}$$

where the unknown constant correlation between the paired data is noted α.

Let $\boldsymbol{A}_i = diag(b''(\theta_{im})a(\phi))$ ($i = 1, \cdots, I$) be the $n_i \times n_i$ matrix corresponding to the variances under the generalized exponential model and $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\alpha)\boldsymbol{A}_i^{1/2}$.

By analogy with Equation C.15, Liang and Zeger (1986) proposed the generalized estimating equations

$$\sum_{i=1}^{I} \boldsymbol{X}_i'\boldsymbol{\Delta}_i\boldsymbol{A}_i\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}. \tag{C.46}$$

These equations do not possess an analytical solution. If the solution exists and is uniquely defined, the equations can be solved by the iterative Newton-Raphson procedure. Moreover, Liang and Zeger (1986) showed that

Theorem C.8.1. *Under mild regularity conditions and given that:*

(i) $\hat{\boldsymbol{\alpha}}$ *is* $I^{\frac{1}{2}}$*-consistent given* β *and* $a(\Phi)$*,*

(ii) $\hat{a}(\Phi)$ *is* $I^{\frac{1}{2}}$*-consistent given* β*,*

(iii) $\partial\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{a}(\Phi))/\partial\boldsymbol{a}(\Phi) \leq \boldsymbol{H}(\boldsymbol{Y}, \boldsymbol{\beta})$ *which is* $O_p(1)$

then $I^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ *is asymptotically multivariate geussian with zero mean and covariance matrix* $\boldsymbol{V}_{\boldsymbol{\beta}}$ *given by*

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \lim_{I\to\infty} I \left[\sum_{i=1}^{I} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right]^{-1} \left[\sum_{i=1}^{I} \boldsymbol{D}_i' \boldsymbol{V}^{-1} cov(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right] \left[\sum_{i=1}^{I} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right]^{-1}$$

where $\boldsymbol{D}_i' = \boldsymbol{X}_i' \boldsymbol{\Delta}_i \boldsymbol{A}_i$*.*

APPENDIX D

Weighted least-squares approach

Weighted least-squares approach is an alternative to the maximum likelihood approach. Let $g = 1, \cdots, G$ index the G conditions under which measurements on the same basic response with K categories are observed. Let $j = 1, \cdots, r$ index the set of categories corresponding to the K^r response profiles associated with the simultaneous classification for the G responses of interest. Let $i = 1, \cdots, s$ index a set of categories corresponding to distinct sub-populations defined in terms of independent variables. If samples of size n_i $(i = 1, \cdots, s)$ are independently selected from the respective sub-populations, the resulting data can be summarized in an $s \times r$ contingency table (Table D.1) where n_{ij} denotes the frequency of profile j in the sample from the ith sub-population.

Table D.1. Observed contingency table

	Response profile categories			
Sub-population	1	...	r	Total
1	n_{11}	...	n_{1r}	n_1
⋮	⋮	⋮	⋮	⋮
s	n_{s1}	...	n_{sr}	n_s

The vector $\boldsymbol{n}_i' = (n_{i1}, \cdots, n_{ir})$ is assumed to follow a multinomial distribution with parameters n_i and $\boldsymbol{\pi}_i' = (\pi_{i1}, \cdots, \pi_{ir})$, where π_{ij} represents the probability that a randomly selected element from the ith population is classified in the jth profile $(\sum_{i=1}^{r} \pi_{ij} = 1$ for $i = 1, \cdots, s)$.

Let $\boldsymbol{p}_i = \boldsymbol{n}_i/n_i$ be the $r \times 1$ vector of observed proportions associated with the sample from the *ith* sub-population and let $\boldsymbol{p}' = (\boldsymbol{p}_1', \cdots, \boldsymbol{p}_s')$. A consistent estimator for the covariance matrix of $\boldsymbol{p}$ is given by the $sr \times sr$ block diagonal matrix $\boldsymbol{V}(\boldsymbol{p})$ with the matrices

$$\boldsymbol{V}_i(\boldsymbol{p}_i) = \frac{1}{n_i}\left[\boldsymbol{D}_{\boldsymbol{p}_i} - \boldsymbol{p}_i\boldsymbol{p}_i'\right] \ (i = 1, \cdots, s) \tag{D.1}$$

on the main diagonal, where $\boldsymbol{D}_{\boldsymbol{p}_i}$ is an $r \times r$ matrix with elements of the vector $\boldsymbol{p}_i$ on the main diagonal.

Let $F_1(\boldsymbol{p}), \cdots, F_u(\boldsymbol{p})$ be a set of u functions of $\boldsymbol{p}$. Each of these functions is assumed to have continuous partial derivatives up to second order with respect to the elements of $\boldsymbol{p}$ within an open region containing $\boldsymbol{\pi} = E(\boldsymbol{p})$. If

$$\boldsymbol{F}' = [\boldsymbol{F}(\boldsymbol{p})]' = [F_1(\boldsymbol{p}), \cdots, F_u(\boldsymbol{p})] \tag{D.2}$$

then a consistent estimator of the covariance matrix of $\boldsymbol{F}$ is the $u \times u$ matrix (Delta method)

$$\boldsymbol{V}_{\boldsymbol{F}} = \boldsymbol{H}\left[\boldsymbol{V}(\boldsymbol{p})\right]\boldsymbol{H}' \tag{D.3}$$

where

$$\boldsymbol{H} = \frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial \boldsymbol{x}}\,|_{\boldsymbol{x}=\boldsymbol{p}}\,. \tag{D.4}$$

It is assumed that the functions comprising $\boldsymbol{F}$ are chosen so that $\boldsymbol{V}_{\boldsymbol{F}}$ is asymptotically nonsingular. The function $\boldsymbol{F}$ is a consistent estimator of $\boldsymbol{F}(\boldsymbol{\pi})$. We assume that the functions $F_1(\boldsymbol{\pi}), \cdots, F_u(\boldsymbol{\pi})$ are jointly independent of one another. The variation among elements of $\boldsymbol{F}(\boldsymbol{\pi})$ can be investigated by fitting linear regression models by the method of weighted least-squares. Assume that

$$\boldsymbol{E}_A\left[\boldsymbol{F}(\boldsymbol{p})\right] = \boldsymbol{F}(\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta} \tag{D.5}$$

where $\boldsymbol{X}$ is a prespecified $u \times p$ design matrix of known coefficients with full rank $p \leq u$, $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of parameters and $\boldsymbol{E}_A$ denotes the '*asymptotic expectation*'.

An appropriate test statistic for the goodness of fit of the model D.5 is

$$Q = Q(\boldsymbol{X}, \boldsymbol{F}) = (\boldsymbol{R}\boldsymbol{F})'\left[\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{F}}\boldsymbol{R}'\right]^{-1}\boldsymbol{R}\boldsymbol{F} \tag{D.6}$$

where $\boldsymbol{R}$ is any full rank $(u-p) \times u$ matrix orthogonal to $\boldsymbol{X}$. Q is approximately distributed according to the chi-square distribution with $(u-p)$ degrees of freedom if the sample sizes n_i are sufficiently 'large' such that the elements of the vector $\boldsymbol{F}$ have an approximately multivariate Normal distribution (Central Limit theorem).

These test statistics, (D.6), are obtained by using weighted least-squares on the basis of the fact that Q is identically equal to

$$Q = (\boldsymbol{F} - \boldsymbol{X}\boldsymbol{b})' \boldsymbol{V}_F^{-1} (\boldsymbol{F} - \boldsymbol{X}\boldsymbol{b}) \tag{D.7}$$

where $\boldsymbol{b} = (\boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{F}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{F}}^{-1} \boldsymbol{F}$ is a BAN estimator for $\boldsymbol{\beta}$. Both Q and $\boldsymbol{b}$ are regarded as having reasonable statistical properties in samples which are sufficiently large for applying CLT to the functions $\boldsymbol{F}$. As a result, a consistent estimator for the covariance matrix of $\boldsymbol{b}$ is given by

$$\boldsymbol{V}_{\boldsymbol{b}} = \left(\boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{F}}^{-1} \boldsymbol{X} \right)^{-1}. \tag{D.8}$$

If the model D.5 does adequately characterize the vector $\boldsymbol{F}(\boldsymbol{\pi})$, tests of linear hypotheses pertaining of the parameters $\boldsymbol{\beta}$ can be undertaken by standard multiple regression procedures. In particular, for a general hypothesis of the form

$$H_0 : \ \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0} \ \text{ versus } H_1 : \ \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{0} \tag{D.9}$$

where $\boldsymbol{C}$ is a known $c \times p$ matrix of full rank $c \leq p$ and $\boldsymbol{0}$ is a $c \times 1$ vector of 0′s, a suitable test statistic is

$$Q_C = (\boldsymbol{C}\boldsymbol{b})' \left[\boldsymbol{C} (\boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{F}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{C}' \right] \boldsymbol{C}\boldsymbol{b} \tag{D.10}$$

which has approximately a chi-square distribution with c degrees of freedom in large samples under H_0 in D.9.

Predicted values for $\boldsymbol{F}(\boldsymbol{\pi})$ based on the model D.5 can be calculated from

$$\widehat{\boldsymbol{F}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X} (\boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{F}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{F}}^{-1} \boldsymbol{F}. \tag{D.11}$$

Thus, consistent estimators for the variances of the elements of $\widehat{\boldsymbol{F}}$ can be obtained from the diagonal elements of

$$\boldsymbol{V}_{\widehat{\boldsymbol{F}}} = \boldsymbol{X} (\boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{F}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}'. \tag{D.12}$$

A wide range of problems in categorical data analysis can be expressed in terms of repeated applications of any sequences of the following matrix operations:

1. linear transformations of the type

$$\boldsymbol{F}_1(\boldsymbol{p}) = \boldsymbol{A}_1 \boldsymbol{p} = \boldsymbol{a}_1$$

2. logarithmic transformations of the type

$$\boldsymbol{F}_2(\boldsymbol{p}) = \ln(\boldsymbol{p}) = \boldsymbol{a}_2$$

3. exponential transformations of the type

$$\boldsymbol{F}_3(\boldsymbol{p}) = \exp(\boldsymbol{p}) = \boldsymbol{a}_3$$

Then the linearized Taylor-series based estimate of the covariance matrix of $\boldsymbol{F}_l$ for $l = 1, 2, 3$ is given by D.3, where the corresponding $\boldsymbol{H}_l$ matrix operator is $\boldsymbol{H}_1 = \boldsymbol{A}_1$, $\boldsymbol{H}_2 = \boldsymbol{D}_{\boldsymbol{p}}^{-1}$ and $\boldsymbol{H}_3 = \boldsymbol{D}_{a_3}$.

Bibliography

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics 44*, 539–548.

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research 1*, 201–218.

Albert, A. and J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*, 1–10.

Banerjee, M., M. Capozzoli, L. McSweeney, and D. Sinha (1993). Beyond kappa: a review of interrater agreement measures. *The Canadian Journal of Statistics 27*, 3–23.

Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates. *Biometrics 52*, 695–702.

Barlow, W., M.-Y. Lai, and S. P. Azen (1991). A comparison of methods for calculating a stratified kappa. *Statistics in Medicine 10*, 1465–1472.

Barnhart, H. X., M. J. Haber, and L. I. Lin (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics 17*, 529–569.

Barnhart, H. X. and J. M. Williamson (2002). Weighted least-squares approach for comparing correlated kappa. *Biometrics 58*, 1012–1019.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports 19*, 3–11.

Basu, S., A. Basu, and A. Raychaudhuri (1999). Measuring agreement between two raters for ordinal response: a model-based approach. *Journal of the Royal Statistical Society, Series D 48*, 339–348.

Becker, M. P. and A. Agresti (1992). Log-linear modelling of pairwise interobserver agreement on a categorical scale. *Statistics in Medicine 11*, 101–114.

Bennett, E. M., R. Alpert, and A. C. Goldstein (1954). Communications through limited response questioning. *The Public Opinion Quarterly, 18*, 303–308.

Bishop, Y. Y. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis. Theory and Practice*. Cambridge: MIT Press.

Blackman, N. J.-M. and J. J. Koval (2000). Interval estimation for cohen's kappa as a measure of agreement. *Statistics in medicine 19*, 723–741.

Bland, A. C., C. D. Kreiter, and J. A. Gordon (2005). The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine 80*, 395–399.

Bland, J. M. and D. G. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet i*, 307–310.

Bloch, D. A. and H. C. Kraemer (1989). 2×2 kappa coefficients: measures of agreement or association. *Biometrics 45*, 269–287.

Brenner, H. and U. Kliebsch (1996). Dependence of weighed kappa coefficients on the number of categories. *Epidemiology 7*, 199–202.

Byrt, T., J. Bishop, and J. B. Carlin (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology 46*, 423–429.

Cantor, A. B. (1996). Sample-size calculations for Cohen's kappa. *Psychological Methods 1*, 150–153.

Carey, V., S. L. Zeger, and P. Diggle (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika 80*, 517–526.

Carrasco, J. L. and L. Jover (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics 59*, 849–858.

Charlin, B., R. Gagnon, L. Sibert, and C. Van der Vleuten (2002). Le test de concordance de script : un instrument d'évaluation du raisonnement clinique. *Pédagogie Médicale 3*, 135–144.

Cicchetti, D. V. and T. Allison (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology 11*, 101–109.

Cicchetti, D. V. and A. R. Feinstein (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology 43*, 551–558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*, 37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin 70*, 213–220.

Collard, A., S. Gelaes, S. Vanbelle, S. Bredart, J.-O. Defraigne, J. Boniver, and B. J.-P. (2009). *Medical Pedagogy To appear*.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin 88*, 322–328.

Coughlin, S. S., L. W. Pickle, M. T. Goodman, and L. R. Wilkens (1992). The logistic modeling of interobserver agreement. *Journal of Clinical Epidemiology 45*, 1237–1241.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika 16*, 297–334.

Davies, M. and J. L. Fleiss (1982). Measuring agreement for multinomial data. *Biometrics 38*, 1047–1051.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology 26*, 297–302.

Donner, A. (1998). Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine 17*, 1157–1168.

Donner, A. and M. Eliasziw (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine 11*, 1511–1519.

Donner, A. and M. Eliasziw (1997). A hierarchical approach to inferences concerning interobserver agreement for multinomial data. *Statistics in medicine 16*, 1097–1106.

Donner, A., M. Eliasziw, and N. Klar (1996). Testing the homogeneity of kappa statistics. *Biometrics 52*, 176–183.

Donner, A. and N. Klar (1996). The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology 49*, 1053–1058.

Donner, A., M. M. Shoukri, N. Klar, and E. Bartfay (2000). Testing the equality of two dependent kappa statistics. *Statistics in Medicine 19*, 373–387.

Eckstein, M. P., T. D. Wickens, G. Aharonov, G. Ruan, C. A. Morioka, and J. S. Whiting (1998). Quantifying the limitations of the use of consensus expert committees in roc studies. Volume 3340, pp. 128–134. SPIE.

Efron, B. and R. J. Tibshirani (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Feigin, P. D. and M. Alvo (1986). Intergroup diversity and concordance for ranking data: an approach via metrics for permutations. *The Annals of Statistics 14*, 691–707.

Feinstein, A. R. and D. V. Cicchetti (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology 43*, 543–549.

Fermanian, J. (1984). Mesure de l'accord entre deux juges. cas qualitatif. *Revue d'Epidémiologie et de Santé Publique 32*, 140–147.

Fisher, R. A. (1958). *Statistical methods for research workers* (13th ed.). New York: Hafner.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*, 378–382.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley.

Fleiss, J. L. and J. Cohen (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability. *Educational and Psychological Measurement 33*, 613–619.

Fleiss, J. L. and J. Cuzick (1979). The reliability of dichotomous judgements: unequal numbers of judges per subject. *Applied Psychological Measurement 3*, 537–542.

Fleiss, J. L. and M. Davies (1982). Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *American Journal of Epidemiology 115*, 841–845.

Fleiss, J. L., J. C. M. Nee, and J. R. Landis (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin 86*, 974–977.

Fleiss, J. L. and P. E. Shrout (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika 43*, 259–262.

Garner, J. B. (1991). The standard error of Cohen's kappa. *Statistics in Medicine 10*, 767–775.

Gilmour, E., T. V. Ellerbrock, J. P. Koulos, M. A. Chiasson, J. Williamson, L. Kuhn, and T. C. J. Wright (1997). Measuring cervical ectopy: direct visual assessment versus computerized planimetry. *American Journal of Obstetrics and Gynecology 176*, 108–111.

Gonin, R., S. R. Lipsitz, G. M. Fitzmaurice, and G. Molenberghs (2000). Regression modelling of weighted κ by using generalized estimating equations. *Journal of the Royal Statistical Society, Series C 49*, 1–18.

Goodman, L. A. and W. H. Kruskal (1954, 1959, 1963, 1972). Measures of association for cross classifications I, II, III, IV. *Journal of the American Statistical Association 49, 54, 58, 67*, 732–764, 123–163, 310–364, 415–421.

Graham, P. (1995). Modelling covariate effects in observer agreement studies: the case of nominal scale agreement. *Statistics in Medicine 14*, 299–310.

Graham, P. and R. Jackson (1993). The analysis of ordinal agreement data: beyond weighted kappa. *Journal of Clinical Epidemiology 46*, 1055–1062.

Grizzle, J. E., C. F. Starmer, and G. G. Koch (1969). Analysis of categorical data by linear models. *Biometrics 25*, 489–504.

Hoehler, F. K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology 53*, 499–503.

Hollander, M. and J. Sethuraman (1978). Testing for agreement between two groups of judges. *Biometrika 65*, 403–411.

Holley, J. W. and J. P. Guilford (1964). A note on the G index of agreement. *Educational and Psychological Measurement 32*, 749–753.

Hosmer, D. and S. Lemeshow (2000). *Applied Logistic Regression* (2nd ed.). New York: John Wiley ans Sons.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin 84*, 289–297.

Kalant, N., M. Berlinguet, J. G. Diodati, L. Dragatakis, and F. Marcotte (2000). How valid are utilization review tools in assessing appropriate use of acute care beds? *Canadian Medical Association Journal 162*, 1809–1813.

King, T. S. and V. M. Chinchilli (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine 20*, 2131–2147.

Klar, N., S. R. Lipsitz, and J. G. Ibrahim (2000). An estimating equations approach for modelling kappa. *Biometrical Journal 42*, 45–58.

Klar, N., S. R. Lipsitz, M. Parzen, and T. Leong (2002). An exact bootstrap interval for κ in small samples. *Journal of the Royal Statistical Society, Series D 51*, 467–478.

Koch, G. G., J. R. Landis, J. L. Freeman, D. H. J. Freeman, and R. G. Lehnen (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics 33*, 133–158.

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika 44*, 461–472.

Kraemer, H. C. (1981). Intergroup concordance: definition and estimation. *Biometrika 68*, 641–646.

Kraemer, H. C. (1992). Measurement of reliability for categorical data in medical research. *Statistical Methods in Medical Research 1*, 183–199.

Kraemer, H. C., S. P. Vyjeyanthi, and A. Noda (2004). Dynamic ambient paradigms. In R. B. D'Agostino (Ed.), *Tutorial in Biostatistics vol 1.*, pp. 85–105. New York: John Wiley and Sons.

Kramer, H. C. (1997). What is the "right" statistical measure of twin concordance (or diagnostic reliability and validity)? *Archives of General Psychiatry 54*, 1121–1124.

Landis, J. R. and G. G. Koch (1975a). A review of statistical methods in the analysis of data arising from observer reliability studies (part I). *Statistica Neerlandica 29*, 101–123.

Landis, J. R. and G. G. Koch (1975b). A review of statistical methods in the analysis of data arising from observer reliability studies (part II). *Statistica Neerlandica 29*, 151–161.

Landis, J. R. and G. G. Koch (1977a). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics 33*, 363–374.

Landis, J. R. and G. G. Koch (1977b). The measurement of observer agreement for categorical data. *Biometrics 33*, 159–174.

Landis, J. R. and G. G. Koch (1977c). A one-way components of variance model for categorical data. *Biometrics 33*, 671–679.

Lantz, C. A. and E. Nebenzahl (1996). Behavior and interpretation of the κ statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology 49*, 431–434.

Lesaffre, E. and A. Albert (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society, Series B 51*, 109–116.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22.

Liang, K.-Y., S. L. Zeger, and B. Qaqish (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B 54*, 3–40.

Light, R. J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin 76*, 365–377.

Lin, H.-M., J. M. Williamson, and S. R. Lipsitz (2003). Calculating power for the comparison of dependent κ-coefficients. *Journal of the Royal Statistical Society, Series C 52*, 391–404.

Lin, L., A. S. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: models, issues and tools. *Journal of American Statistical Association 97*, 257–270.

Lin, L., A. S. Hedayat, and W. Wu (2007). A unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics 17*, 629–652.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics 45*, 255–268.

Lipsitz, S. R. and G. M. Fitzmaurice (1996). Esimating equations for measures of association between repeated binary responses. *Biometrics 52*, 903–912.

Lipsitz, S. R., N. M. Laird, and D. P. Harrington (1990). Maximum likelihood regression methods for paired binary data. *Statistics in Medicine 9*, 1517–1525.

Lipsitz, S. R., M. Parzen, G. M. Fitzmaurice, and N. Klar (2003). A two-stage logistic regression model for analyzing inter-rater agreement. *Psychometrika 68*, 289–298.

Lipsitz, S. R., J. Williamson, N. Klar, J. Ibrahim, and M. Parzen (2001). A simple method for estimating a regression model for κ between a pair of raters. *Journal of the Royal Statistical Society, Series A 164*, 449–465.

Lord, F. M. and M. R. Novick (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clinical and Experimental Pharmacology and Physiology 29*, 527–536.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *The British Journal of Psychiatry 130*, 79–83.

Mc Cullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B 42*, 109–142.

McGraw, K. O. and S. P. Wong (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods 1*, 30–46.

McKenzie, D. P., A. J. Mackinnon, N. Peladeau, P. Onghena, P. C. Bruce, D. M. Clarke, S. Harrigan, and P. D. McGorry (1996). Comparing correlated kappas by resampling: is one level of agreement significantly different from another? *Journal of Psychiatric Research 30*, 483–492.

Mielke, P. W. and K. J. Berry (2008). Resampling probability values for weighted kappa with multiple raters. *Psychological Reports 102*, 606–613.

Miller, D. P., K. F. O'Shaughnessy, S. A. Wood, and R. A. Castellino (2004). Gold standards and expert panels: a pulmonary nodule case study with challenges and solutions. Volume 5372, pp. 173–184. SPIE.

Miller, M. E., C. S. Davis, and J. R. Landis (1993). The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. *Biometrics 49*, 1033–1044.

Nam, J.-M. (2000). Interval estimation of the kappa coefficient with binary classification an equal marginal probability model. *Biometrics 56*, 583–585.

Nam, J.-M. (2003). Homogeneity score test for the intraclass version of the kappa statistics and sample-size determination in multiple or stratified studies. *Biometrics 59*, 1027–1035.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A 135*, 370–384.

Oden, N. L. (1991). Estimating kappa from binocular data. *Statistics in Medicine 10*, 1303–1311.

Olkin, I. and J. W. Pratt (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics 29*, 201–211.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A 195*, 1–45.

Perkins, S. M. and M. P. Becker (2002). Assessing rater agreement using marginal association models. *Statistics in Medicine 21*, 1743–1760.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics 44*, 1033–1048.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika 43*, 353–360.

Raine, R., C. Sanderson, A. Hutchings, S. Carter, K. Larkin, and N. Black (2004). An experimental study of determinants of group judgments in clinical guideline development. *Lancet 364*, 429–437.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.). New York: Wiley.

Roberts, C. and R. McNamee (1998). A matrix of kappa-type coefficients to assess the reliability of nominal scales. *Statistics in Medicine 17*, 471–488.

Roberts, C. and R. McNamee (2005). Assessing the reliability of ordered categorical scales using kappa-type statistics. *Statistical Methods in Medical Research 14*, 493–514.

Rogot, E. and I. D. Goldberg (1966). A proposed index for measuring agreement in test-retest studies. *Journal of chronic diseases 19*, 991–1006.

Ruperto, N., A. Ravelli, S. Oliveira, M. Alessio, D. Mihaylova, S. Pasic, E. Cortis, M. Apaz, R. Burgos-Vargas, F. Kanakoudi-Tsakalidou, X. Norambuena, F. Corona, V. Gerloni, S. Hagelberg, A. Aggarwal, P. Dolezalova, C. M. Saad, S.-C. Bae, R. Vesely, T. Avcin, H. Foster, C. Duarte, T. Herlin, G. Horneff, L. Lepore, M. van Rossum, L. Trail, A. Pistorio, B. Andersson-Gare, E. H. Giannini, A. Martini, PEDIATRIC RHEUMATOLOGY INTERNATIONAL TRIALS ORGANIZATION (PRINTO), and PEDIATRIC RHEUMATOLOGY COLLABORATIVE STUDY GROUP (PRCSG) (2006). The pediatric rheumatology international trials organization/american college of rheumatology provisional criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: Prospective validation of the definition of improvement. *Arthritis Care and Research 55*, 355–363.

Salerno, S. M., P. C. Alguire, and S. W. Waxman (2003). Competency in interpretation of 12-lead electrocardiograms: A summary and appraisal of published evidence. *Annals of Internal Medicine 138*, 751–760.

Satterwhaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin 2*, 110–114.

Schouten, H. J. A. (1982). Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica 36*, 45–61.

Schouten, H. J. A. (1993). Estimating kappa from binocular data and comparing marginal probabilities. *Statistics in Medicine 12*, 2207–2217.

Schucany, W. R. and W. H. Frawley (1973). A rank test for two group concordance. *Psychometrika 38*, 249–258.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology 55*, 289–303.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relation to other rater agreement statistics for metric scales. *Educational and Psychological Measurement 64*, 243–253.

Schuster, C. and D. A. Smith (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods 7*, 384–395.

Schuster, C. and D. A. Smith (2006). Estimating with a latent class model the reliability of nominal judgments upon which two raters agree. *Educational and Psychological Measurement 66*, 739–747.

Schuster, C. S. and D. A. Smith (2005). Dispersion-weighted kappa: an integrative framework for metric and nominal scale agreement coefficients. *Psychometrika 70*, 135–146.

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly 19*, 321–325.

Shoukri, M. M. (2004). *Measures of interobserver agreement*. Boca Raton: Chapman and Hall/CRC.

Shoukri, M. M., S. W. Martin, and I. U. H. Mian (1995). Maximum likelihood estimation of the kappa coefficient from models of matched binary responses. *Statistics in Medicine 14*, 83–99.

Shoukri, M. M. and I. U. H. Mian (1996). Maximum likelihood estimation of the kappa coefficient from bivariate logistic regression. *Statistics in Medicine 15*, 1409–1419.

Shrout, P. E. and J. L. Fleiss (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin 86*, 420–428.

Smith, R., A. J. Copas, M. Prince, B. George, A. S. Walker, and S. T. Sadiq (2003). Poor sensitivity and consistency of microscopy in the diagnosis of low grade non-gonococcal urethritis. *Sexually Transmitted Infections 79*, 487–490.

Soeken, K. L. and P. A. Prescott (1986). Issues in the use of kappa to estimate reliability. *Medical care 24*, 733–741.

Strike, P. W. (1991). *Statistical Methods in Laboratory Medicine*. Oxford: Butterworth-Heinemann.

Tanner, M. A. and M. A. Young (1985a). Modeling agreement among raters. *Journal of the American Statistical Association 80*, 175–180.

Tanner, M. A. and M. A. Young (1985b). Modeling ordinal scale disagreement. *Psychological Bulletin 98*, 408–415.

Thompson, W. D. and S. D. Walter (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology 41*, 949–958.

Thomson, J. R. (2001). Estimating equations for kappa statistics. *Statistics in Medicine 20*, 2895–2906.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples (abstract). *Annals of Mathematical Statistics 29*, 614.

Uebersax, J. S. (1988). Validity inferences from interobserver agreement. *Psychological Bulletin 104*, 405–416.

Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology 58*, 655–661.

van Hoeij, M. J. W., J. C. W. Haarhuis, R. F. A. Wierstra, and P. van Beukelen (2004). Developing a classification tool based on Bloom's taxonomy to assess the cognitive level of short essay questions. *Journal of Veterinary Medical Education 31*, 261–267.

Vanbelle, S. (2002). Accord entre observateurs et coéfficient kappa de Cohen. Master's thesis, University of Liège, Belgium.

Vanbelle, S. and A. Albert (2008). A boostrap method for comparing correlated kappa coefficients. *Journal of Statistical Computation and Simulation 78*, 1009–1015.

Vanbelle, S. and A. Albert (2009a). Agreement between an isolated rater and a group of raters. *Statistica Neerlandica 1*, 82–100.

Vanbelle, S. and A. Albert (2009b). Agreement between two groups of raters. *Psychometrika*.

Vanbelle, S. and A. Albert (2009c). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology 6*, 157–163.

Vanbelle, S., V. Massart, D. Giet, and A. Albert (2007). Test de concordance de script: un nouveau mode d'établissement des scores limitant l'effet du hasard. *Pédagogie Médicale 8*, 71–81.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika 63*, 27–32.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

Williams, G. W. (1976). Comparing the joint agreement of several raters with another rater. *Biometrics 32*, 619–627.

Williamson, J. M. and A. K. Manatunga (1997). Assessing interrater agreement from dependent data. *Biometrics 53*, 707–714.

Williamson, J. M., A. K. Manatunga, and S. R. Lipsitz (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics 1*, 191–202.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin 103*, 374–378.