

Kai Bruchlos

Inter-Rater Reliability: Chance-corrected Measures

THM-Hochschulschriften Band 22

Kai Bruchlos

Inter-Rater Reliability: Chance-corrected
Measures

THM-Hochschulschriften Band 22

THM-Hochschulschriften Band 22

© 2022 Kai Bruchlos

Technische Hochschule Mittelhessen

Fachbereich Mathematik, Naturwissenschaften und
Datenverarbeitung

Herausgeber der THM-Hochschulschriften:

Der Präsident der Technischen Hochschule Mittelhessen

Alle Rechte vorbehalten, Nachdruck, auch auszugsweise, nur mit
schriftlicher Genehmigung und Quellenangabe.

Die Hochschulschriften sind online abrufbar:

www.thm.de/bibliothek/thm-hochschulschriften

ISSN (Print) 2568-0846

ISSN (Online) 2568-3020

Inter-Rater Reliability: Chance-corrected Measures

Kai Bruchlos

Abstract

The assessment of the inter-rater reliability requires the reduction of the observed agreement by chance agreement. There are several measures for this purpose. We consider some measures for nominal variables and investigate the mathematical characteristics of these measures on the basis of two mathematical models. Moreover we check if they are chance-corrected measures. Furthermore we introduce qualitative features for measures and check them on the measures. Finally we also give guidelines for the interpretation of the values of a measure and consider chance agreement as random measurement error.

Keywords: agreement for nominal categories, chance-corrected measure, inter-rater agreement, inter-rater reliability, kappa statistic

MSC Class: 62H20; 62P10; 62P15; 62P25

1 Introduction

We consider a scientific investigation where observers (raters, judges) have to classify objects (individuals, things) into categories. The prerequisite is that the classification of the objects is independent of the selected observers. The observers should be interchangeably (see Gwet 2014, p. 4) which means the results are reproducible (see Cohen 1960, p. 37). Inter-rater reliability is the degree of agreement among observers in such an investigation. There are many measures to calculate the degree (see Banerjee et al. 1999). An important property for a measure is “freedom from random measurement error” (Schinka and Velicer 2003, p. 399).

There are several reasons for disagreement among observers. These reasons could have been personal preferences, different interpretations of categories, uncertainty about the correct category or misunderstanding about categorization (see Gwet 2014, p. 11, 29). Therefore the observed agreement

(sample agreement) may be partially or completely random. We remove the chance from the observed agreement with the objective to obtain the chance-corrected agreement (see Gwet 2014, p. 15).

The experiment (survey) we are looking at has the following statistical properties: 1. N units (objects), 2. nominal variable with $k > 1$ values (categories), 3. two observers operating independently.

The paper is organized as follows. Firstly section 2 presents the basic characteristics of the observed agreement and the chance-corrected agreement. Secondly section 3 introduces two mathematical models for estimating and assessing measures. Thirdly in section 4 we consider the measure of Cohen, Scott, Gwet and Brennan, Prediger. Finally section 5 concludes the paper with a discussion of the features of the measures and clues for the application.

2 Properties of the observed agreement

We start with a contingency table for two observers A and B, k categories, N observations (units) and n_{ij} number of observations that observer A and B respectively classified into category i and j respectively:

		Observer B			
		1	2	...	k
Observer A	1	n_{11}	n_{12}	...	n_{1k}
	2	n_{21}	n_{22}	...	n_{2k}
	⋮	⋮	⋮		⋮
	k	n_{k1}	n_{k2}	...	n_{kk}

n_{ii} is the number of observed agreements in the category i , $i = 1, \dots, k$. We have

$$N = \sum_{i=1}^k \sum_{j=1}^k n_{ij} .$$

We consider the relative observed agreement among observers

$$p_0 := \frac{1}{N} \cdot \sum_{i=1}^k n_{ii} .$$

The range of p_0 is $0 \leq p_0 \leq 1$.

Now suppose that we only know the classification in categories by observer A and B. The n_{ij} are not available. Which maximum value can p_0 achieve? The contingency table must be switched from absolute frequency

to relative frequency to answer this question:

		Observer B				
Categories		1	2	...	k	marginal totals
Observer A	1	f_{11}	f_{12}	...	f_{1k}	$f_{1\cdot}$
	2	f_{21}	f_{22}	...	f_{2k}	$f_{2\cdot}$
	\vdots	\vdots	\vdots		\vdots	\vdots
	k	f_{k1}	f_{k2}	...	f_{kk}	$f_{k\cdot}$
	marginal totals	$f_{\cdot 1}$	$f_{\cdot 2}$...	$f_{\cdot k}$	

Here

$$f_{ij} := \frac{n_{ij}}{N}, \quad i, j = 1, \dots, k$$

is the relative frequency and

$$f_{i\cdot} := \sum_{j=1}^k f_{ij}, \quad i = 1, \dots, k, \quad f_{\cdot j} := \sum_{i=1}^k f_{ij}, \quad j = 1, \dots, k$$

are the marginal totals. It applies

$$1 = \sum_{i=1}^k \sum_{j=1}^k f_{ij} = \sum_{i=1}^k f_{i\cdot} = \sum_{j=1}^k f_{\cdot j} .$$

Proposition 1 *If only the marginal totals are given, then the maximal possible value of p_0 is*

$$\max(p_0) = \sum_{i=1}^k \min(f_{i\cdot}, f_{\cdot i}) .$$

Proof: Let a_1, \dots, a_k be the result of the classification by observer A and b_1, \dots, b_k the result of the classification by observer B. The maximum value for n_{ii} is $\min(a_i, b_i)$. It follows:

$$\max(p_0) = \frac{1}{N} \cdot \sum_{i=1}^k \min(a_i, b_i) = \sum_{i=1}^k \min(f_{i\cdot}, f_{\cdot i})$$

□

Corollary 1 $\max(p_0) = 1$ if and only if $f_i = f_{\cdot i}$ for all $i = 1, \dots, k$.

Proof: Suppose that $\max(p_0) = 1$. Suppose furthermore there exist $i \in \{1, \dots, k\}$ such that $f_i < f_{\cdot i}$. Then we have

$$1 = \max(p_0) = \sum_{i=1}^k \min(f_i, f_{\cdot i}) < \sum_{i=1}^k f_i = 1 .$$

Contradiction. The case $f_{\cdot i} > f_i$ is similar. Conversely, if $f_{\cdot i} = f_i$ for all $i = 1, \dots, k$, then $\max(p_0) = \sum_{i=1}^k f_i = 1$.

□

We need the following approach for chance-corrected measures:

$$\kappa_c := \frac{n_{11} + \dots + n_{kk} - c_{11} - \dots - c_{kk}}{N - c_{11} - \dots - c_{kk}} = \frac{\sum_{i=1}^k f_{ii} - \frac{1}{N} \cdot \sum_{i=1}^k c_{ii}}{1 - \frac{1}{N} \cdot \sum_{i=1}^k c_{ii}}$$

c_{ii} is the number of random matches of the category i , $i = 1, \dots, k$. The denominator is the reference value and has to be reduced by the number of random matches accordingly. However, this approach has a structural weakness:

Proposition 2 *Let the value of $c_{11} + \dots + c_{kk}$ be fixed. The higher the value of $n_{11} + \dots + n_{kk}$, the smaller the influence of $c_{11} + \dots + c_{kk}$ on κ_c is.*

Proof: The statement follows from

$$\kappa_c = 1 - \frac{N - (n_{11} + \dots + n_{kk})}{N - (c_{11} + \dots + c_{kk})}.$$

□

Example 1 *Let $N = 10$ and $c_{11} + \dots + c_{kk} = 3$. For $n_{11} + \dots + n_{kk} = 5$ we get $\kappa_c = 0,29$ and for $n_{11} + \dots + n_{kk} = 9$ we get $\kappa_c = 0,86$. In the second case, the value of $n_{11} + \dots + n_{kk}$ is by a factor of 1.8 higher, while the value of κ_c is by a factor of 3.*

The main question is how to estimate the agreement by chance (see Gwet 2014, p. 34 et seq.), strictly speaking the number of random matches c_{ii} . We need probabilistic models for the different estimation methods and the interpretation of chance.

3 Probabilistic Models

We use two models. Model 1 is the conventional urn model. (see Heumann et al. 2016, p. 97 et seq.) There are k balls in the urn labeled with the numbers 1 to k . We randomly draw one ball with replacement two times and with consideration of the order of the balls. We get the probability space (see Pestman 1998, p. 13)

$$(\{1, \dots, k\} \times \{1, \dots, k\}, \mathcal{P}(\{1, \dots, k\} \times \{1, \dots, k\}), P),$$

where P is the product measure $P = P_1 \otimes P_2$ with $P(A \times B) = P_1(A) \cdot P_2(B)$ for all $A, B \subset \{1, \dots, k\}$ and

$$P_1(i) = P_2(i) = \frac{1}{k} \quad \text{for all } i = 1, \dots, k .$$

It follows $P(i, i) = \frac{1}{k^2}$ for $i = 1, \dots, k$.

For model 2, (X, Y) is a two-dimensional real random variable, F_{XY} the distribution function of X and Y , F_X the distribution function of X and F_Y the distribution function of Y . (see Pestman 1998, p. 37, 40.) The feature that the two observers are operating independently means X and Y are stochastically independent:

$$F_{XY}(x, y) = F_X(x) \cdot F_Y(y) .$$

In this context, F_X and F_Y are called marginal distribution functions of (X, Y) (see Fisz 1976, p. 63 et seq.; Pestman 1998, p. 13 et seq.). If the random variables X and Y have probability density functions f_X and f_Y the probability density functions of (X, Y) is given by (Pestman 1998, p. 15, Theorem I.3.2)

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y) .$$

f_X and f_Y are called marginal density function of (X, Y) .

We can use the last equation to estimate the agreements by chance for observer A and B:

$$\frac{c_{ii}}{N} := f_{AB}(i, i) = f_A(i) \cdot f_B(i)$$

So the estimation of the agreement by chance is determined by the marginal density functions in model 2.

4 Estimating the agreement by chance

In this section we take a look at some measures that are supposed to be chance-corrected.

4.1 Cohen

Under model 2, Cohen 1960 estimates the marginal density function f_A and f_B respectively with the marginal totals $f_{i\cdot}$ and $f_{\cdot j}$ respectively. However, this choice results in structural problems.

Cohen's agreement by chance is

$$p_c := \sum_{i=1}^k f_{i\cdot} \cdot f_{\cdot i} = \sum_{i=1}^k \hat{f}_A(i) \cdot \hat{f}_B(i) = \sum_{i=1}^k \hat{f}_{AB}(i, i) = \sum_{i=1}^k \frac{\widehat{c_{ii}}}{N} .$$

So we have

$$\kappa := \kappa_c(p_c) = \frac{p_0 - p_c}{1 - p_c} .$$

The range of p_c is $0 \leq p_c \leq 1$. – Now we want to study properties of κ . It follows from Corollary 1

Lemma 1 *It applies $\max(\kappa) = 1$ for all k .*

Lemma 2 (Cohen 1960) *If only the marginal totals are given, then the maximal and minimal respectively possible value of κ is*

$$\min(\kappa) = -\frac{p_c}{1-p_c} \quad \text{and} \quad \max(\kappa) = \frac{\max(p_0) - p_c}{1-p_c} \quad \text{respectively.}$$

In particular $-\infty < \kappa \leq 1$.

Proposition 3 *If $\kappa = 1$, then the marginal totals are identical: $f_i = f_{\cdot i}$ for all $i = 1, \dots, k$.*

Proof: From $\kappa = 1$ it follows

$$1 = \kappa = \frac{1 - p_c}{1 - p_c} = \frac{\max(p_0) - p_c}{1 - p_c}.$$

The statement follows from Corollary 1.

□

Corollary 2 *$\kappa = 1$ and complete agreement ($\sum n_{ii} = N$) are equivalent.*

What if the value of κ means in the case of no complete agreement for example $\sum n_{ii} = N - 1$? The answer to this question is difficult:

Proposition 4 *Let the value of $p_0 < 1$ be fixed. The smaller the value of p_c , the higher the value of κ is.*

Proof: The statement follows from

$$\kappa = 1 - \frac{1 - p_0}{1 - p_c}.$$

□

See also Feinstein and Cicchetti 1990, page 544.

Furthermore, we consider the effects of marginal totals on κ of answering the question. We follow the argumentation of Feinstein and Cicchetti 1990. They investigate 2×2 contingency tables to study the effects well:

		Observer B		marginal totals
		1	2	
Observer A	1	n_{11}	n_{12}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$n_{2\cdot}$
marginal totals		$n_{\cdot 1}$	$n_{\cdot 2}$	

They use the proportion of marginal totals:

$$v := \frac{n_{.1}}{N} \quad \text{and} \quad w := \frac{n_{1.}}{N}$$

Therefore we have

$$n_{.2} = (1 - v) \cdot N, \quad n_{2.} = (1 - w) \cdot N \quad \text{and} \quad p_c = 2vw - v - w + 1 .$$

Nomenclature 1 (i) Marginal totals are called **balanced**, if $v, w \approx 0.5$.

(ii) Marginal totals are called **symmetrical**, if $v \approx w$.

(iii) Marginal totals are called **asymmetrical**, if $v \approx 1 - w$.

Corollary 3 Let the marginal totals be symmetrical. All other things being equal, κ has a lower value in the situation of not balanced marginal totals than in the situation of balanced marginal totals.

Proof: We have to show $p_c(v, v) \geq p_c(0.5, 0.5)$. It is $p_c(v, v) = 2v^2 - 2v + 1$ and $p_c(0.5, 0.5) = 0.5$. The statement follows from

$$0 \leq v(1 - v) \leq 0.25 \Rightarrow -0.25 \leq v(v - 1) \leq 0 \Rightarrow -0.5 \leq 2v^2 - 2v \leq 0 .$$

□

Corollary 4 All other things being equal, κ has a lower value in the situation of perfect symmetrical marginal totals ($v = w > 0.5$) than in the situation of less-than-perfect symmetrical marginal totals ($v, v - \varepsilon$ and $w - \varepsilon, w$ respectively, $\varepsilon > 0$).

Proof: Let $\tilde{w} := v - \varepsilon, 0 < \varepsilon < v \leq 1$ and $\tilde{v} := w - \varepsilon, 0 < \varepsilon < w \leq 1$. We have to show $p_c(v, w) > p_c(v, \tilde{w})$ and $p_c(v, w) > p_c(\tilde{v}, w)$.

We are looking at the case $p_c(v, w) > p_c(v, \tilde{w})$. We have $v + v > 1 \Rightarrow 2v\varepsilon > \varepsilon$ and therefore

$$p_c(v, w) = 2v^2 - 2v + 1 > 2v^2 - 2v\varepsilon - 2v + \varepsilon + 1 = p_c(v, \tilde{w}) .$$

The case $p_c(v, w) > p_c(\tilde{v}, w)$ is similar.

□

Remark 1 The statement of Corollary 4 also applies in similar situations. An example can be found in Feinstein and Cicchetti 1990, page 546.

Corollary 5 Let $v, w \neq 0.5$ and $\tilde{v} := 1 - v, \tilde{w} := 1 - w$. All other things being equal, κ has a higher value in the situation of asymmetrical marginal totals (v, \tilde{w} and \tilde{v}, w) than in the situation of perfect symmetrical marginal totals ($v = w$).

Proof: We have to show $p_c(v, w) > p_c(v, \tilde{w})$ and $p_c(v, w) > p_c(\tilde{v}, w)$. We are looking at the case $p_c(v, w) > p_c(v, \tilde{w})$. It is $p_c(v, w) = 2(v^2 - v) + 1$ and $p_c(v, \tilde{w}) = 2(v - v^2)$. We have

$$v(1 - v) < 0.25 \Rightarrow 2(v - v^2) < 1 - 2(v - v^2) .$$

This shows the statement. The case $p_c(v, w) > p_c(\tilde{v}, w)$ is similar. □

Remark 2 (i) Corollary 5 also holds if $v \approx w, \tilde{w} \approx 1 - w$ or $\tilde{v} \approx 1 - v$. Feinstein and Cicchetti 1990 give an example of this on page 545.

(ii) The condition “ $v > 0.5$ or $\tilde{w} > 0.5$ and $\tilde{v} > 0.5$ or $w > 0.5$ respectively” is easy to fulfill. Where appropriate, we define $v := n_2/N$ and $w := n_2./N$.

As the above Corollaries suggest, the point $v = w = 0.5$ has a specific relevance:

Proposition 5 $p_c(v, w)$ has no local extreme value, but a saddle point at $v = w = 0.5$.

Proof: We determine the first partial derivatives of $p_c(v, w)$:

$$\frac{\partial}{\partial v} p_c(v, w) = 2w - 1, \quad \frac{\partial}{\partial w} p_c(v, w) = 2v - 1$$

The condition $\text{grad } p_c(v, w) = (0, 0)$ leads to $0 = 2w - 1, 0 = 2v - 1$. We have the critical point $(0.5, 0.5)$. Second partial derivative test:

$$\frac{\partial^2}{\partial v^2} p_c(v, w) = 0, \quad \frac{\partial^2}{\partial w^2} p_c(v, w) = 0, \quad \frac{\partial^2}{\partial v \partial w} p_c(v, w) = 2$$

Since $D(0.5, 0.5) = -4$, $p_c(v, w)$ has no local extreme value, but a saddle point at $(0.5, 0.5)$. □

4.2 Scott and Gwet

Under model 2, Scott 1955 considers like Cohen 1960 the marginal totals $f_{i \cdot}$ and $f_{\cdot j}$ respectively as realisations of the marginal density function f_A and f_B respectively. The advanced part is that he stabilizes the results by averaging:

$$\pi_c := \sum_{i=1}^k \frac{f_{i \cdot} + f_{\cdot i}}{2} \cdot \frac{f_{i \cdot} + f_{\cdot i}}{2} = \sum_{i=1}^k \hat{f}_A(i) \cdot \hat{f}_B(i) = \sum_{i=1}^k \hat{f}_{AB}(i, i) = \sum_{i=1}^k \frac{\widehat{c_{ii}}}{N} .$$

He defines

$$\pi := \kappa_c \left(\frac{f_{i\cdot} + f_{\cdot i}}{2} \right) = \frac{p_0 - \pi_c}{1 - \pi_c} .$$

Under model 2, Gwet 2014, p. 104 considers the marginal totals $f_{i\cdot}$ as realisations of the marginal density function f_A and defines the estimator for the marginal density function f_B as a weighted complement:

$$\begin{aligned} \hat{\gamma}_c &:= \frac{1}{k-1} \cdot \sum_{i=1}^k \frac{f_{i\cdot} + f_{\cdot i}}{2} \cdot \left(1 - \frac{f_{i\cdot} + f_{\cdot i}}{2} \right) = \sum_{i=1}^k \hat{f}_A(i) \cdot \hat{f}_B(i) \\ &= \sum_{i=1}^k \hat{f}_{AB}(i, i) = \sum_{i=1}^k \frac{\widehat{C_{ii}}}{N} . \end{aligned}$$

So we have the coefficient

$$\hat{\gamma}_1 := \kappa_c(\hat{\gamma}_c) = \frac{p_0 - \hat{\gamma}_c}{1 - \hat{\gamma}_c} .$$

Lemma 1 and 2, Proposition 3, Corollary 2 and Proposition 4 also apply to π and $\hat{\gamma}_1$.

Proposition 4 is the basis for the structural influence (balanced, symmetrical, asymmetrical) of the marginal totals on Cohen's κ . Since π_c and $\hat{\gamma}_c$ are also calculated with the marginal totals, there is a similar structural influence for π and $\hat{\gamma}_1$.

4.3 Brennan, Prediger

Under model 2, Brennan and Prediger 1981 choose on pages 692 et seq. the discrete uniform distribution for the marginal density functions. So the agreement by chance is

$$\sum_{i=1}^k \frac{1}{k} \cdot \frac{1}{k} = \sum_{i=1}^k \hat{f}_A(i) \cdot \hat{f}_B(i) = \sum_{i=1}^k \hat{f}_{AB}(i, i) = \sum_{i=1}^k \frac{\widehat{C_{ii}}}{N} = \frac{1}{k} .$$

They define

$$\kappa_k := \kappa_c \left(\frac{1}{k} \right) = \frac{p_0 - \frac{1}{k}}{1 - \frac{1}{k}} \in \left[-\frac{1}{k-1}; 1 \right] .$$

Guttman 1946, has already chosen this approach.

What properties does κ_k have? From Corollary 1 it follows

Lemma 3 *It applies $\max(\kappa_k) = 1$ for all k .*

If the design of the experiment is established, then the marginal totals have the same effect on κ_k for every observer. Indeed the number of categories has an effect on κ_k :

Lemma 4 *It applies for fixed p_0*

$$\lim_{k \rightarrow \infty} \kappa_k = p_0 .$$

In other words, the higher the number of categories, the smaller the value of agreement by chance is. However that is only one side of the coin.

As the number of categories increases, the agreement is decreasing (see Bennett et al. 1954, p. 306). This is because it is more difficult for observers to differentiate between categories as numbers increase. If we view $p_0(k) := p_0$ as a realisation of a random variable Z_k , then the increasing insecurity in decision for the appropriate category can be expressed as follows:

$$\lim_{k \rightarrow \infty} \tilde{P}(Z_k > \varepsilon) = 0$$

Z_k converges in probability towards 0. Here Z_2, Z_3, \dots are random variables on a probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$.

In summary, as the number of categories increases, $\frac{1}{k}$ and p_0 become smaller. This shows the influence of the number of categories on the experiment.

5 Discussion

In addition to the purpose to remove the chance from the observed agreement, the measure should have the following features. The values of the measure enable the comparison of couples of observers in the same experiment and also the comparison of experiments. Furthermore the values of the measure should be meaningful. Which measure has these features?

First of all, there are limits to the meaningfulness by Proposition 2. This means that the influence of chance on all measures of the form κ_c including the measures considered here depends on the magnitude of the observed agreement. Moreover, if agreement by chance is estimated by a sample, the characteristics of the observers are taken into account. Thus, the meaningfulness is further limited. The last argument doesn't apply to κ_k .

Until now, p_0 has been considered as observed agreement. Under model 2, p_0 can also be considered as special sum of stochastic independence. This applies equally to p_c . Accordingly, $p_0 - p_c$ is the difference between part of the observed stochastic independence and the corresponding part of the estimated theoretical stochastic independence (see Pestman 1998, p. 172 et seq.; Heumann et al. 2016, p. 238). So, why is p_0 reduced by chance? After all, the quantities for the deviation of observed stochastic independence and theoretical stochastic independence in the χ^2 -test on statistical independence are the differences $f_{ij} - f_i \cdot f_j$ (see Gwet 2014, p. 35).

Furthermore, the effects of marginal totals on κ don't enable the comparison of couples of observers in the same experiment or the comparison of experiments. Moreover, it is difficult to interpret the magnitude of κ . (see Gwet 2014, p. 34, 166; Landis and Koch 1977, p. 164 et seq; Stoyan et al. 2018.)

Similar arguments apply to Scott's π and Gwet's $\hat{\gamma}_1$.

Brennan and Prediger choose the same marginal density function for all observers. So we always have the same effects of marginal totals and we can compare couples of observers in the same experiment. Experiments are comparably, if the experiments have the same number of categories.

Under model 2, $p_0 - \frac{1}{k}$ can be considered as difference between part of the observed stochastic independence and the corresponding part of estimated theoretical stochastic independence. This applies only if the marginal density function of the observers is the discrete uniform density function. Under model 1, $p_0 - \frac{1}{k}$ is the difference between the observed agreement and the sum of random choices of a category:

$$\sum_{i=1}^k \frac{\widehat{c_{ii}}}{N} = \frac{1}{k} = \sum_{i=1}^k \frac{1}{k^2} = \sum_{i=1}^k P(i, i)$$

If the discrete uniform distribution describes chance, then κ_k is a chance-corrected measure.

What about the significance of the values of κ_k ? In the case of negative values, it is very unlikely that the observers agree. Corollary 2 shows that $\kappa_k = 1$ and complete agreement are equivalent. An agreement should be likely if p_0 is at least twice the size of the random agreement. For $p_0 = \frac{2}{k}$, κ_k has the value

$$\frac{\frac{2}{k} - \frac{1}{k}}{1 - \frac{1}{k}} = \frac{\frac{1}{k}}{\frac{k-1}{k}} = \frac{1}{k-1} .$$

So if the values are in the interval $[0; \frac{1}{k-1})$, then an agreement is unlikely. With increasing k the fraction $\frac{1}{k-1}$ becomes smaller and smaller. Analogously the observed agreement gets smaller and smaller. A statement about values in the interval $[\frac{1}{k-1}; 1)$ depends on the design of the experiment. In this context it is also important that the categories are clearly different.

Now we would like to enquire the question of which measure is chance-corrected. Gwet 2014, p. 32 writes about it: "The idea of adjusting the percent agreement p_0 for chance agreement is often controversial, and the definition of what constitutes chance agreement is part of the problem." (see Brennan and Prediger 1981, p. 688 et seq.) We understand by chance agreement the random measurement error. When measuring physical quantities, it is common to assume the normal distribution for the random measurement error (see Pestman 1998, p. 62). Which distribution for the random measurement error is suitable here?

Ten timekeeper take the time of a sprinter. Some measured times will be above the true value, some below. The smaller the distance of the measured value from the true value, the more likely it is. So we assume the normal distribution for the random measurement error. Let's move on to our case, the nominal scale. Ten people are to assign a brush to a living area: Bath, kitchen, living room, workroom or bedroom. The brush belongs in the kitchen. What is the most probable random false classification now?

In the case of timekeepers, we use the total ordering of the real numbers to determine how far away a false classification is from the correct one. We then determine that the closer the false classification is, the more likely it is. Distance matrix, similarity measure and probability measure correspond.

Every colour can be assigned to a specific wavelength. If the categories are colors, then there is a natural order. We have a ordinal variable. A corresponding probability measure can be defined: If green is the right color, then the false classification of yellow is more likely than red. This probability measure can be used as the random measurement error for all observers.

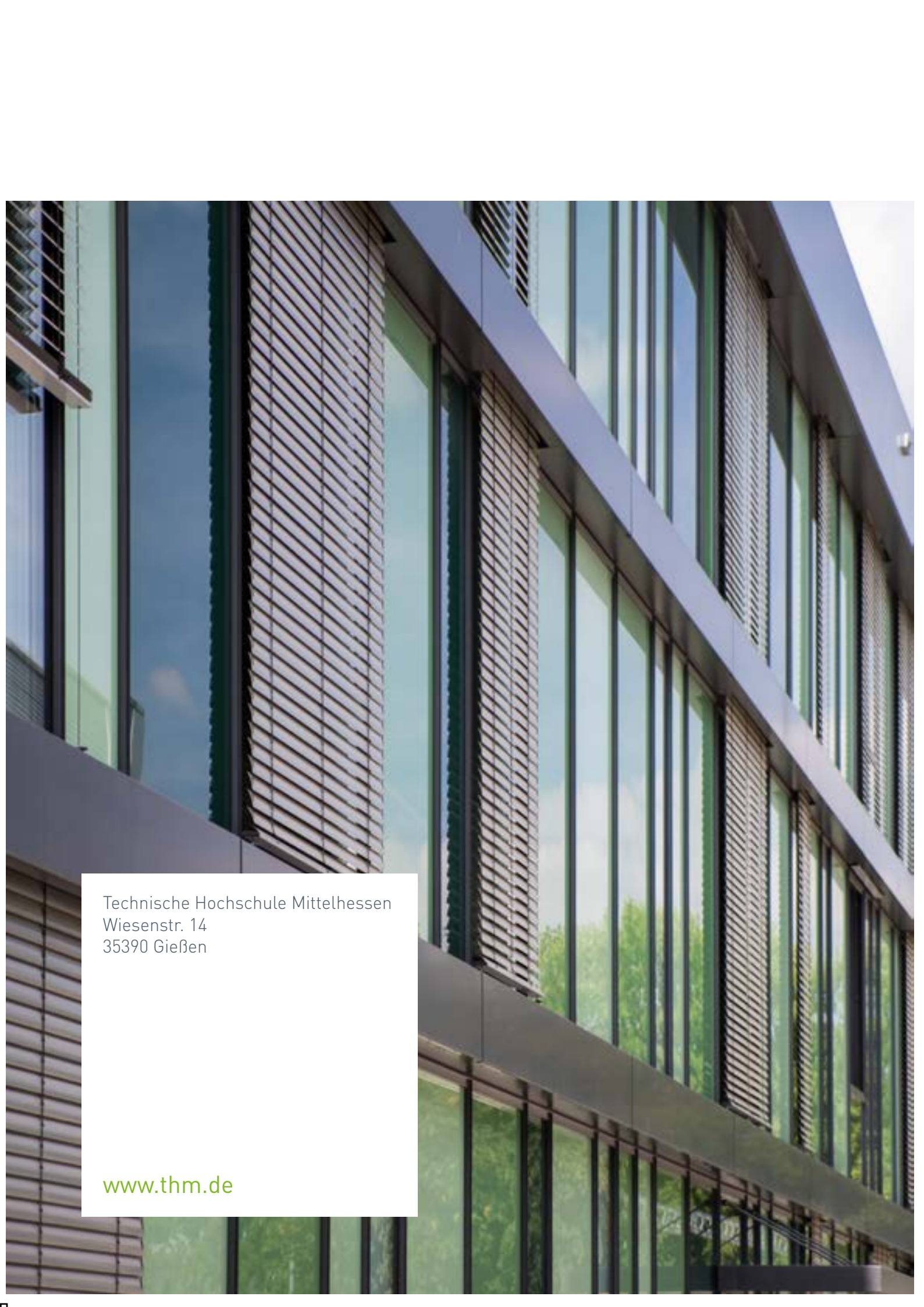
What probability measure should be chosen if there is no natural order or similarity measure, as in the example of the living area? Any false classifications are equally likely. The random false classification of the brush to the living room is just as likely as the random false classification to the workroom. It's like the urn model, model 1. The corresponding probability measure is the discrete uniform distribution.

In summary, assuming that chance agreement is the random measurement error, we can state: In order to avoid difficulties with different marginal totals, a marginal density function for the random measurement error is chosen for all observers. If there is a natural order, a distance matrix or a similarity measure, the corresponding probability measure determines the marginal density function. Otherwise, the marginal distribution is equal to the discrete uniform distribution, that is the approach of Brennan and Prediger.

References

- Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D.: Beyond kappa: A review of interrater agreement measures. *Can. J. Stat.* 27, 3–23 (1999)
- Bennett, E.M., Alpert, R., Goldstein, A.C.: Communications through Limited-Response Questioning. *Public Opin. Q.* 18(3), 303–308 (1954)
- Brennan, R.L., Prediger, D.J.: Coefficient kappa: Some uses, misuses, and alternatives. *Educ. Psychol. Meas.* 41, 687–699 (1981)
- Cohen, J.: A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46 (1960)
- Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* 43, 543–549 (1990)

- Fisz, M.: Probability Theory and Mathematical Statistics, 3rd edn. Wiley, New York (1976)
- Guttman, L.: The test-retest reliability of qualitative data. *Psychometrika* 11(2), 81–94 (1946)
- Gwet, K.: Handbook of Inter-Rater Reliability, 4th edn. Advanced Analytics, Gaithersburg (2014)
- Heumann, C., Schomaker, M., Shalabh: Introduction to Statistics and Data Analysis. Springer Switzerland, Cham (2016)
- Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* 33(1), 159–174 (1977)
- Pestman, W.: Mathematical Statistics. De Gruyter, Berlin (1998)
- Scott, W.A.: Reliability of content analysis: the case of nominal scale coding. *Public Opin. Q.* 19(3), 321–325 (1955)
- Schinka, J.A., Velicer, W.F. (eds.): Research Methods in Psychology. In: Weiner IB (ed.) *Handbook of Psychology*, Vol. 2. Wiley, Hoboken NJ (2003)
- Stoyan, D., Pommerening, A., Hummel, M., Kopp-Schneider, A.: Multiple-rater kappas for binary data: Models and interpretation. *Biom. J.* 60(2), 381–394 (2018)

A photograph of a modern building facade featuring large glass windows and metal panels. The windows reflect the sky and surrounding greenery. The building has a clean, industrial aesthetic with dark frames and light-colored metal accents.

Technische Hochschule Mittelhessen
Wiesenstr. 14
35390 Gießen

www.thm.de