Statistical Models for the Precision of Categorical Measurement
van Wieringen, W.N.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
2 The assessment of precision of binary measurement systems

In this chapter we study how to assess precision of a binary measurement system. Evidently, model (1.1) of the continuous case is not applicable in the situation of binary measurement. We propose to model the outcome of a binary measurement system analysis experiment by a latent class model. Next, we relate the model parameters to the concept of measurement precision. We use the latent class model to evaluate alternative approaches to evaluate the measurement system, namely the kappa statistic, the intraclass correlation coefficient and log-linear models. This comparison sheds light upon what is the best method to analyze a measurement system analysis study for binary measurements. We conclude with an example illustrating all techniques discussed. This chapter is based on Van Wieringen and Van den Heuvel (2003).

2.1 Latent class model

Consider a rater measuring an object with a binary measurement system. This measurement $X$ will be either zero or one, and is taken to be a random variable that is Bernoulli distributed with parameter $p = P(X = 1)$. We assume that the reference value of the measured object is also either zero or one. The reference value of an object, henceforth called $Y$, is also taken to be Bernoulli distributed with parameter $\theta = P(Y = 1)$, the probability that an object is of good quality.

The measurement $X$ is dependent on $Y$, the reference value of the measured object. We define $\pi(y) = P(X = 1|Y = y)$, the conditional probability of an object being measured as one given the reference value $Y$. The unconditional probability that a randomly selected object is measured as $x \in \{0,1\}$ is:

$$P(X = x) = P(X = x|Y = 0)P(Y = 0) + P(X = x|Y = 1)P(Y = 1)$$

$$= (1 - \theta)\pi(0)^x(1 - \pi(0))^{(1-x)} + \theta\pi(1)^x(1 - \pi(1))^{(1-x)}.$$  \hspace{1cm} (2.1)

For the situation involving multiple raters we have visualized this measurement process in figure 2.1.
The assessment of precision of binary measurement systems

The outcome of the measurement system analysis experiment is modelled by a latent class model which specifies the joint probability distribution of the set of rater responses. Latent class analysis distinguishes between a manifest variable (the measurement of a rater) and an unobserved, latent variable (the reference value of the object). The latter is used to explain the correlation structure in the (observed) former. Crucial to this approach is that it assumes conditional independence. That is, given the latent variable, the manifest variables are independent of each another. Conditional independence can be formulated as:

$$P(X_1, X_2, \ldots, X_m|Y) = \prod_{j=1}^{m} P(X_j|Y),$$

(2.2)

i.e., given the reference value of the object, the raters $j = 1, \ldots, m$ measure independently.

Since both the observed and latent variable are Bernoulli, the unconditional probability that rater $j$ measures an object $i = 1, \ldots, n$ as good can be written as in (2.1). Using this and (2.2) we specify the model underlying the latent class analysis. Let $X$ denote the $n \times m$ matrix containing the data from the measurement system analysis experiment, with $X_{ij}$ the measurement of rater $j$ of object $i$, given by:

$$X = \begin{pmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nm} \end{pmatrix}$$
The likelihood function of the joint response of the raters of the sample, \( X \), is:
\[
L(X; \Psi) = \prod_{i=1}^{n} \left( (1 - \theta) \prod_{j=1}^{m} (\pi_j(0))^{X_{ij}} (1 - \pi_j(0))^{1-X_{ij}} + \theta \prod_{j=1}^{m} (\pi_j(1))^{X_{ij}} (1 - \pi_j(1))^{1-X_{ij}} \right),
\]
where we substituted \( P(Y_i = 1) = \theta \) and \( P(X_{ij} = 1|Y_i = y) = \pi_j(y) \) for all \( i \) and \( j \), and \( \Psi = (\theta, \pi_1(1), \ldots, \pi_m(1), \pi_1(0), \ldots, \pi_m(0)) \). To ensure identifiability of the model additional restrictions need to be imposed. In the particular case where each rater makes only one measurement, at least 3 raters need to be involved and it is required that \( \theta \in (0, 1) \) and \( 1 \geq \pi_j(1) > \pi_j(0) \geq 0 \) for all \( j \). Restrictions for the general case and the proof that they guarantee identifiability are given in the next chapter.

Equation (2.3) plus the additional restrictions enable one to use a maximum likelihood procedure to estimate the parameters (Bartholomew and Knott, 1999; Boyles, 2001). To find a maximum likelihood estimate for \( \Psi \), instead of applying the Newton-Raphson algorithm, the E-M algorithm is used. It has been shown that the sequence of estimates produced by the E-M algorithm converges to a maximum of the likelihood function (McLachlan and Krishnan, 1997). This is also described in the next chapter.

### 2.1.1 Latent class method

The precision of a measurement system is assessed on the basis of an experiment. The design of the experiment should allow for the estimation of all parameters in the model. A balanced design, where all objects of the sample are measured under all circumstances of the factors under study, repetitively, meets this requirement. We restrict ourselves to one factor, which we take to be the raters executing the measurement. For this purpose \( n \) objects are selected randomly, and are measured by all \( m \) raters. The outcome of this experiment can be described by the latent class model and all its parameters can be estimated.

Besides describing the outcome of the experiment, the latent class method enables a natural operationalization of the precision of the measurement. The only measurement error in the case of binary measurements is that of misclassification. Therefore, for binary measurement an operational definition of precision should be related to the probability of misclassification. However, the probability of misclassification itself depends on the quality of the measured objects, whereas the evaluation of the measurement system is preferably independent of the quality of the measured objects. Therefore, we adopt from Uebersax (1988) the terms sensitivity and specificity. Sensitivity is defined as \( \pi_m(1) = P(X = 1|Y = 1) \), the probability that a good object is measured as such. Specificity is defined as \( 1 - \pi_m(0) = P(X = 0|Y = 0) \), the probability that a bad object is measured as bad. Sensitivity \( \pi_m(1) \) and specificity \( 1 - \pi_m(0) \) are related to the type I error and type II error as \( 1 - \pi_m(1) \) and \( \pi_m(0) \), respectively.

This operationalization allows – given the process parameter \( \theta \) and estimates \( \hat{\pi}_1(0), \ldots, \hat{\pi}_m(0) \) and \( \hat{\pi}_1(1), \ldots, \hat{\pi}_m(1) \) for the parameters – calculation of the probability of misclassification. Assume for simplicity that all raters measure an equal share of the objects. Then, for any quality \( \theta \) of the sample, the estimated probability of misclassification is:
\[
P(\text{misclassification}) = \frac{1}{m} \sum_{j=1}^{m} (\theta(1 - \hat{\pi}_j(1)) + (1 - \theta)\hat{\pi}_j(0)).
\]
If raters measure unequal shares, small modifications are required. In addition, for a particular object one can indicate which category the object is most likely to originate from: category \( y \) that maximizes: 
\[ P(Y = y | X_1, X_2, \ldots, X_m). \]

## 2.2 Alternative methods

### 2.2.1 Measure of agreement based on kappa

Many measures representing the quality of binary measurement systems have been proposed and can be found in Goodman and Kruskal (1954) and the review papers of Landis and Koch (1975a,b). Cohen (1960) introduces a measure of agreement called the kappa. This statistic has been proposed as a statistic for the evaluation of categorical measurement systems, confer Dunn (1989), Futrell (1995) and AIAG (2002).

A concept related to precision in the context of binary measurement is agreement. Two measurements of one object agree if they are identical. Agreement is measured by the \( \kappa \) statistic, which represents the degree of agreement between two raters, based on how they classify a sample of objects into a number of categories. However, some agreement may be due to chance. The \( \kappa \) statistic, corrected for agreement by chance and normalized, is of the form:

\[
\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{2.5}
\]

Here \( P_o \) is the observed proportion of agreement and \( P_e \) the expected proportion of agreement due to chance. The \( \kappa \) statistic attains the value 1 when there is perfect agreement, 0 if observed agreement is merely due to chance and negative values when the amount of agreement is less than is to be expected on the basis of chance. Frequently the observed proportion is used to evaluate the measurement process. However, \( P_o \) confounds systematic agreement with agreement by chance, whereas \( \kappa \) focusses on systematic agreement only.

As a comparison consider a multiple choice exam. Marks are calculated in accordance with (2.5). That is, the proportion of questions the examinee answered correctly, \( P_o \), is lessened by the expected proportion of questions he would have answered correctly had he chosen his answers randomly, \( P_e \). This difference is scaled to translate it into a mark.

Cohen (1960) specifies, for any pair of raters \( j_1, j_2 \), the terms in (2.5) as

\[
P_o = \sum_{x=0}^{1} p_{j_1,j_2}(x, x) \quad \text{and} \quad P_e = \sum_{x=0}^{1} p_{j_1}(x) p_{j_2}(x).
\]

Here \( P_o \) is the proportion of objects with matching measurements of raters \( j_1 \) and \( j_2 \) and \( p_{j_1,j_2}(x, x) \) denotes the proportion of objects that have been measured as \( x \) by raters \( j_1 \) and \( j_2 \). The expected proportion of agreement \( P_e \) is based on the individual marginal distributions of each rater. The marginal proportion for rater \( j \) and category \( x \) is denoted by \( p_j(x) \). Thus, in line with the traditional contingency table setting Cohen (1960) observes that in the situation where measurements are made completely random the responses of the raters are independent.

Due to the way \( P_e \) is calculated, \( \kappa \) may give values that are counter-intuitive. For instance, suppose that all raters measure almost all objects in the same category (small object variation). Then, \( \kappa \) is small, as \( P_e \) is large. Thus, \( \kappa \) confounds to some extent precision of the measurement system with object variation. Similarly, let one rater measure almost all objects in one category, and the other rater almost all of them in a different category (systematic rater difference). Then,
2.2 Alternative methods

$P_e$ approaches its minimum and causes a relatively high $\kappa$. Thus, whereas $\kappa$ is designed to measure systematic rater differences, it ignores them to some extent. These are called the paradoxes of the kappa (Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990). In this context it has been argued (see Brennan and Prediger, 1981) to define agreement by chance as completely random, i.e., the raters assign the objects to any category with equal probability.

Landis and Koch (1977) proposes the following table which expresses the relationship between the value of $\kappa$ and the corresponding evaluation of the measurement system. The authors suggest that this classification is arbitrary.

<table>
<thead>
<tr>
<th>Quality of measurements</th>
<th>Kappa value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>&lt; 0.00</td>
</tr>
<tr>
<td>Slight</td>
<td>0.00–0.20</td>
</tr>
<tr>
<td>Fair</td>
<td>0.21–0.40</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.41–0.60</td>
</tr>
<tr>
<td>Substantial</td>
<td>0.61–0.80</td>
</tr>
<tr>
<td>Almost perfect</td>
<td>0.81–1.00</td>
</tr>
</tbody>
</table>

Table 2.1: Correspondence between $\kappa$ and the quality of measurements

Another approach is to test $H_0 : \kappa = 0$ against $H_A : \kappa \neq 0$, thus testing whether agreement is substantial, or merely due to chance. However, this approach changes the question from “How good is the measurement process?”, to “Do we have a measurement process at all?”. For more on test procedures and moments of the $\kappa$ see Everitt (1968) and Hubert (1977).

2.2.2 Kappa for multiple raters

We point out briefly how kappa extends to the situation of more than two raters. Since at least two raters are required for agreement, Fleiss (1971) suggests that the degree of agreement may be expressed in terms of the proportion of agreeing pairs. If there are $m$ raters, the maximum number of agreeing pairs possible per object equals $\frac{1}{2}m(m - 1)$. To estimate the proportion of agreeing pairs per object, Fleiss proposes the sum of the number of agreeing pairs per category:

$$ P_o = \frac{1}{nm(m - 1)} \left( \sum_{i=1}^{n} \sum_{x=0}^{1} Z_i(x)(Z_i(x) - 1) \right), $$

with $Z_i(x) = \sum_{j=1}^{m} #(X_{ij} = x)$ the number of times object $i$ has been classified as $x$. The expected proportion of agreement is given by:

$$ P_e = \frac{2}{m(m - 1)} \sum_{j_1 < j_2} \sum_{x=0}^{1} p_{j_1}(x) p_{j_2}(x). $$

Each pair of raters enters the sum only once. $P_e$ estimates, under the assumption of independence, the probability that two randomly selected raters classify an object into the same category, based on the individual marginal proportions of the raters. We have adopted Conger (1980) here instead of Fleiss (1971), with the main difference that Conger allows the raters to have different marginal distributions, and calculate $P_e$ without rater replacement. This has
the advantage that it is conceptually in line with Cohen (1960). This is illustrated by the fact that \( \kappa \) for multiple raters equals the average of all pairwise \( \kappa \)'s, if either there is independence among all raters or their marginal probabilities are equal. One may generalize this approach by considering the other tuples of agreeing raters.

### 2.2.3 Kappa statistic from the perspective of the latent class model

Using the latent class model, we rewrite \( \kappa \) in terms of the parameters of the latent class model. We limit ourselves to two raters\(^\dagger\), to avoid cumbersome notational issues. The observed agreement is the probability that both raters make the same measurement:

\[
P_o = P(X_1 = X_2) = \sum_{x,y=0}^1 |1 - y - \theta| (x - \pi_1(y)) (x - \pi_2(y)).
\]

The expected proportion of agreement, i.e., the probability that by chance the raters measure identically, is given by

\[
P_e = \sum_{x=0}^1 P(X_1 = x) P(X_2 = x) = \sum_{x=0}^1 p_1(x) p_2(x),
\]

with \( p_j(x) = P(X_j = x) \) defined analogous to (2.1). Reformulating (2.5) in terms of the latent class parameters yields a \( \kappa \) that depends on the \( \pi_j(y) \) and \( \theta \). This is displayed graphically, for an arbitrary choice of the \( \pi_j(y) \), by plotting \( \kappa \) against \( \theta \) (see figure 2.2). Thus, for a single measurement system \( \kappa \) can differ substantially from one measurement system analysis experiment to another depending on the quality of the measured objects.

Given that kappa depends on the process parameter \( \theta \), one may argue that the criteria on the kappa, as proposed in Landis and Koch (1977) should be adjusted accordingly, confer Elffers (2001). As the criteria themselves are arbitrary so will be their adjustments.

\(^\dagger\)This violates the identification restrictions, but can be coped with by requiring, in addition to \( \pi_j(1) > \pi_j(0) \), that \( \pi_j(1) = 1 - \pi_j(0) \) for all \( j \).
2.2 Alternative methods

The latent class model is a model for the outcome of a measurement system analysis experiment, whereas \( \kappa \) is trying to summarize all aspects of a measurement system into one number. When \( \kappa \) indicates that the measurement system is not up to standard, it provides no clues how this has arisen. The estimated latent class model, on the other hand, yields information about the individual rater performances, thus giving insight in how discrepancies between measurements have come about.

2.2.4 Intraclass correlation coefficient

The social sciences interpret precision as reliability, which is the consistency with which a measurement system measures a certain property, or, equivalently, the correlation between multiple measurements of the same object. Reliability is often expressed in the form of an intraclass correlation coefficient (Lord and Novick, 1968; Shrout and Fleisch, 1979).

Let \( X_{ij} \) be the measurement of an arbitrary object \( i \) by rater \( j \). Again, it is assumed that \( X_{ij} \) is Bernoulli distributed with parameters \( p_j = P(X_{ij} = 1) \) for all \( i \). For binary measurements the intraclass correlation coefficient is called the \( \phi \) coefficient and (for two raters) defined as

\[
\phi = \frac{\text{Cov}(X_{i1}, X_{i2})}{\sqrt{\text{Var}(X_{i1}) \cdot \text{Var}(X_{i2})}} = \frac{P(X_{i1} = 1, X_{i2} = 1) - p_1 p_2}{\sqrt{p_1 (1 - p_1) p_2 (1 - p_2)}}.
\]

As other product moment correlation coefficients \( \phi \) only assumes values in the interval [-1,1].

The \( \phi \) coefficient is estimated by replacing all the terms in the righthand side of (2.6) by their corresponding estimates: \( \hat{p}_1 = \frac{1}{n} \sum_{i=1}^{n} X_{i1}, \hat{p}_2 = \frac{1}{n} \sum_{i=1}^{n} X_{i2}, \) and \( P(X_{i1} = 1, X_{i2} = 1) \) by \( \frac{1}{n} \sum_{i=1}^{n} X_{i1} X_{i2} \).

For the situation involving \( m > 2 \) raters Fleiss (1965) and Bartko and Carpenter (1976) propose to evaluate the reliability by means of the average of the \( \phi \) coefficients of all possible rater pairs, where they assume that \( p_j = p \) for \( j = 1, \ldots, m \). The \( \phi \) coefficient for multiple raters is then estimated by \( \phi = (P - p^2)/(p - p^2) \) where

\[
p = \frac{1}{n m} \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij} \quad \text{and} \quad P = \frac{2}{n m (m - 1)} \sum_{i=1}^{n} \sum_{j_1=1}^{m-1} \sum_{j_2=j_1+1}^{m} X_{ij_1} X_{ij_2}
\]

When using the intraclass correlation coefficient as the statistic representing the quality of measurements, from Wheeler and Lyday (1989) one can deduce the criteria in table 2.2. The

<table>
<thead>
<tr>
<th>Criterion</th>
<th>( \phi )</th>
<th>&lt; 0.60</th>
<th>0.60-0.90</th>
<th>0.90-1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of measurements</td>
<td>Inadequate</td>
<td>Moderate</td>
<td>Adequate</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.2:** Correspondence between \( \phi \) and the quality of measurements

criteria in table 2.2 apply to intraclass correlation coefficients for continuous measurements. We assume they can be used for \( \phi \) coefficient.
2.2.5 The intraclass correlation coefficient from the perspective of the latent class model

As with the kappa statistic we use the latent class model to study the intraclass correlation coefficient, and restrict the comparison to the two raters case. The numerator of $\phi$ becomes:

$$
\text{Cov}(X_{i1}, X_{i2}) = \sum_{x_1, x_2=0}^{1} (x_1 - p_1)(x_2 - p_2)P(X_{i1} = x_1, X_{i2} = x_2)
$$

$$
= \sum_{x_1, x_2=0}^{1} \left( (x_1 - \theta \pi_1(1) - (1 - \theta) \pi_1(0)) \times (x_2 - \theta \pi_2(1) - (1 - \theta) \pi_2(0)) \times \left( \theta \pi_1(1)^x_1 \left(1 - \pi_1(1)\right)^{1-x_1} \pi_2(1)^x_2 \left(1 - \pi_2(1)\right)^{1-x_2} + (1 - \theta) \pi_1(0)^x_1 \left(1 - \pi_1(0)\right)^{1-x_1} \pi_2(0)^x_2 \left(1 - \pi_2(0)\right)^{1-x_2} \right) \right)
$$

$$
= \theta (1 - \theta) (\pi_1(1) - \pi_1(0))(\pi_2(1) - \pi_2(0)),
$$

and for its denominator $\sqrt{\text{Var}(X_{i1}) \cdot \text{Var}(X_{i2})}$ we have:

$$
\text{Var}(X_{ij}) = \sum_{x=0}^{1} (x - p_j)^2 P(X_{ij} = x) = p_j - p_j^2 \quad j = 1, 2.
$$

As the formula for $\phi$ is not a transparent expression, we resort to visual means to illustrate the relation between $\phi$ and the parameters of the latent class model. The surface in figure 2.3.a represents $\phi$ against $\pi_1(1)$ and $\pi_2(1)$ (to obtain a 3-dimensional graph we have fixed $\theta$ and taken $\pi_j(0) = 1 - \pi_j(1)$ for all $j$). This corresponds with the intuitive idea of $\phi$: $\phi = 1$ if the raters measure similarly (i.e. $\pi_j(1)$ equals 1 for all $m$) and $\phi = 0$ if the raters both rate randomly (i.e. $\pi_1(1) = \frac{1}{2} = \pi_2(1)$).

**Figure 2.3.a:** $\phi$ vs. $\pi_1(1)$ and $\pi_2(1)$

**Figure 2.3.b:** $\phi$ vs. $\theta$

Figure 2.3.b shows that $\phi$ (like $\kappa$) also depends on $\theta$. The $\phi$ statistic thus evaluates the measurement system in relation to the process (with parameter $\theta$). As $\theta$ may vary from process
2.2 Alternative methods

to process the evaluation does not apply to other processes. Moreover, \( \phi \) is not capable of evaluating a measurement system independent of the process. For a comparison consider the Gauge R&R statistic mentioned in chapter 1. The Gauge R&R statistic is not independent of the process as it involves the process spread. To evaluate the measurement system independent of the process one would only use the measurement spread.

Like \( \kappa \), \( \phi \) is a summary statistic, providing only aggregated information, which is of limited use when the measurement system needs improvement.

2.2.6 Log-linear model

Like the kappa method, Tanner and Young (1985) interprets precision as agreement. Instead of defining a measure for agreement, they model agreement. They use the rater measurements to construct a contingency table. The cells of this table are modelled by a log-linear model with two components: one representing the effect of chance, and the other representing the effect of rater agreement.

Let \( X_i = (X_{i1}, X_{i2}, \ldots, X_{im}) \) be the measurements of the \( m \) raters on object \( i \). For each \( m \)-tuple \( \mathbf{x} = (x_1, x_2, \ldots, x_m) \) with \( x_j \in \{0, 1\} \), define \( n(\mathbf{x}) = \sum_{i=1}^{n} \#(X_i = \mathbf{x}) \). \( n(\mathbf{x}) \) is the number of times \( m \)-tuple \( \mathbf{x} \) appears in the measurement system analysis experiment. Tanner and Young assume that \( n(\mathbf{x}) \) is strictly positive. This is a remarkable assumption. When dealing with a precise measurement system, one expects to find (mainly) the patterns \( \mathbf{x} = (0, 0, \ldots, 0) \) and \( \mathbf{x} = (1, 1, \ldots, 1) \). Therefore, one would expect patterns for which \( n(\mathbf{x}) \) equals 0.

Tanner and Young consider the \( n(\mathbf{x}) \) as the cells of a contingency table. Table 2.3 visualizes this for two raters. The main diagonal cells of the contingency table represent the agreement between the raters. Tanner and Young study agreement by comparing the frequencies in these diagonal cells to the expected cell count under an independence model (i.e., all raters measure independently and their marginal distributions yield the expected number of times \( \mathbf{x} \) will occur).

Conventionally, contingency tables are modelled by log-linear models. Therefore, the independence model is given by:

\[
\ln (E_n(\mathbf{x})) = u + \sum_{j=1}^{m} u_j(x_j). \quad (2.7)
\]

Tanner and Young call \( u \) the overall effect and \( u_j(x_j) \) the effect of category \( x_j \) of the \( j \)-th rater. Model (2.7) is an alternative way of stating that the cell counts are explained by the marginal proportions of the raters. From this perspective \( u_j(x_j) \) can be viewed as the difference between
the proportion of rater $j$ measuring an object as $x_j$ and the overall proportion. Added to model (2.7) should be the restriction

$$
\sum_{x_j=0}^1 u_j(x_j) = 0 \quad \text{for all } j. \quad (2.8)
$$

This makes model (2.7) identifiable and assures that the marginal proportions sum to 1 for each rater.

A second term is added to model (2.7), which accounts for the discrepancies between the observed and expected cell counts of the diagonal cells:

$$
\ln (E(n(x))) = u + \sum_{j=1}^m u_j(x_j) + \delta(x), \quad (2.9)
$$

with

$$
\delta(x) = \begin{cases} 
  c & \text{if } x \text{ is a diagonal cell} \\
  0 & \text{otherwise,}
\end{cases}
$$

where $c$ is a constant that reduces the discrepancy between the observed and the expected cell count of the diagonal cells. Tanner and Young interpret $c$ as the effect due to agreement among the raters.

Estimates of the parameters are obtained by a maximum likelihood procedure, where it is assumed that the contingency table can be described by a multinomial distribution. A significant discrepancy between the observed diagonal cells and their expected cell count under the independence model corresponds to the significance of the agreement. The significance of the discrepancy (and thus of the agreement) is assessed by testing whether model (2.9) fits the data significantly better than model (2.7).

Evaluating the measurement system by testing the significance of agreement is in fact answering the question “Do we have a measurement system at all?” Moreover, it is not clear how significance of agreement relates to the consequences, e.g., the number of incorrectly measured objects, of the use of a measurement system.

### 2.2.7 The log-linear model from the perspective of the latent class model

For the comparison between the log-linear model approach and the latent class model we limit ourselves – as before – to the two raters case. The log-linear model is:

$$
\ln(E(n(x))) = u + (-1)^x_1 u_1 + (-1)^x_2 u_2 + \delta(x), \quad (2.10)
$$

where

$$
\delta(x) = \begin{cases} 
  c & \text{if } x \text{ is a diagonal cell} \\
  0 & \text{otherwise.}
\end{cases}
$$

To see what the model actually describes, we have rewritten the agreement contribution $c$ in terms of the latent class parameters:

$$
c = \frac{1}{4} \ln \left( \frac{E(n(0,0)) \cdot E(n(1,1))}{E(n(0,1)) \cdot E(n(1,0))} \right), \quad (2.11)
$$
with
\[ E(n(x)) = n \left( \theta \prod_{j=1}^{2} |x_j - \pi_j(1)| + (1 - \theta) \prod_{j=1}^{2} |x_j - \pi_j(0)| \right). \]

Model (2.10) is saturated, therefore we can write \( c \) as an explicit expression by solving the model in terms of the expected cell counts.

Plotting \( c \) against \( \theta \) (see figure 2.4) reveals they are related (where the \( \pi_m(i) \) are fixed as for the kappa in figure 2.2). This implies that the evaluation of a measurement system by means of

![Figure 2.4: Agreement contribution \( c \) against \( \theta \)](image)

the log-linear model method is not independent of the process parameter \( \theta \).

### 2.3 Example

At an engine manufacturer components are examined on dirt, for too much dirt may cause an engine to break down. For the purpose of examination a tape is affixed to the component. The tape is detached and placed under a microscope, magnified thirty times and photographed. The photograph is compared with a number of references, covering all the varieties of contamination. These references are divided into two categories, one representing the acceptable (clean) surfaces and the other the unacceptable (contaminated) surfaces. A rater decides which reference the photograph resembles best, indirectly judging whether the component is suitable for production or needs to be cleaned first.

To assess the quality of the measurement system we have set up an experiment where three raters measured 20 objects, according to the procedure described above, in random order. Per component only one tape is gathered, which is measured by all raters. The data have been reproduced in table 2.4. We illustrate the methods described in this chapter by applying them to the described measurement system for dirt on engine components.

Using the E-M algorithm as in McLachlan and Krishnan (1997), which maximizes the likelihood function (2.3), we find \( \hat{\theta} = 0.13 \), the sensitivity of each rater \( \hat{\pi}_A(1) = 0.99, \hat{\pi}_B(1) = 0.99, \hat{\pi}_C(1) = 0.89 \), and the specificity for each rater \( 1 - \hat{\pi}_A(0) = 0.58, 1 - \hat{\pi}_B(0) = 0.80 \) and \( 1 - \hat{\pi}_C(0) = 0.50 \). These estimates show the individual rater performances. All raters are good at judging a good object as such. Raters \( A \) and \( C \) have a tendency to mistake bad objects for
The assessment of precision of binary measurement systems

### Experimental data

<table>
<thead>
<tr>
<th>Object</th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater C</th>
<th>Total Good</th>
<th>Total Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>6</td>
<td>11</td>
<td>27</td>
<td>33</td>
</tr>
</tbody>
</table>

**Table 2.4:** Data, with 1 = good and 0 = bad

Good objects, giving too optimistic an impression of the sample. This means, if raters A, B and C measure the engine components, we substitute the found estimates into

\[
P(\text{misclassification}) = \frac{1}{m} \sum_{j=1}^{m} \left( \hat{\theta}(1 - \hat{\pi}_j(1)) + (1 - \hat{\theta})\hat{\pi}_j(0) \right),
\]

we have a probability of 0.33 on a wrongly measured object (cf. formula (2.4)). In table 2.5 the expected frequencies according to the latent class model (LCM) for each response pattern are given. They hardly deviate from the observed frequencies.

Table 2.5 displays the expected frequencies according to the log-linear agreement model, including \( c \), estimated as \( \hat{c} = 0.63 \). When we incorporate this term in the model the Pearson \( \chi^2 \) goodness-of-fit statistic changes from 0.58 to 0.79. Neither one exceeds the \( \alpha \)-level of 0.05.

The intraclass correlation coefficient and kappa for the three raters combined and each possible pair are given in table 2.6. All these indices have a large deviation from their ideal value. This provides no information on the individual rater level to see who needs attention in the improvement process. This is partially due to the fact that the raters have not measured objects repetitively. If the raters measure the objects repetitively \( k \) and \( \phi \) can be calculated for the raters individually. This enables the evaluation of the consistency of each rater.
2.4 Conclusion

In the literature, no truly satisfactory approach for measurement system analysis was found for binary measurement, despite the fact that binary measurements are often encountered in practice. For measurement system analysis experiments with binary measurements we adopt the design used in the continuous setting: each rater involved in the experiment measures all selected objects, preferably repetitively. We introduced the latent class model to model the outcome of such an experiment. This model involves several parameters that all have a clear interpretation. Furthermore, in the paradigm of this model we gave an operational definition for the measurement precision sensible to binary measurements, and directly related to the parameters of the model. Once all parameters are estimated, we have a clear insight into the consequences of applying a measurement system. This serves as the basis for the evaluation of the measurement system.

A comparison of the latent class method to alternative approaches leads to the conclusion that the former has some considerable advantages:

- A one dimensional index (kappa, κ, phi, φ, and delta, δ, in the log-linear model approach)
gives far less insight in the merits of a measurement system then the latent class model
with its $2m$ estimated parameters $\pi_1(1), \ldots, \pi_m(1), \pi_1(0), \ldots, \pi_m(0)$.

$\kappa, \phi$ and $\delta$ evaluate a measurement system in relation to the quality of the process in
which it is to be applied. This is reflected in the dependency of all three statistics on $\theta$,
which represents the quality of the process. In itself it is useful to evaluate a measurement
system in relation to its application, but more insight is gained from splitting this issue
into two sub-issues:

1. How good is the measurement system itself (independent of its application)?
2. Is this good enough for the intended application?

The latent class model answers the first issue in the form of the sensitivity and specificity.
Consideration of the consequences of a misclassification, combined with their estimated
frequency, provide the information to decide in the second issue.

In standard measurement system analysis experiments objects are measured more than
once. It is hard to see how the methods deal with repeated measurements of the objects. Only
the latent class model has a natural extension to experiments with repeated measurements.