TECHNICAL REPORT

ISO/TR 27877

First edition
2021-10

Statistical analysis for evaluating the precision of binary measurement methods and their results

Analyse statistique pour l'évaluation de la fidélité des méthodes de mesure binaire et de leurs résultats

Reference number
ISO/TR 27877:2021(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword ..... **v**

Introduction ..... **vi**

1 Scope **1**

2 Normative references **1**

3 Terms and definitions, and symbols **1**
3.1 Terms and definitions 1
3.2 Symbols 3

4 Overview **4**

5 Examples used in this document **6**
5.1 Case 1: Listeria monocytogenes 6
5.2 Case 2: Human cell line activation test (h-CLAT)-1 6
5.3 Case 3: Intraracheal administration testing 7
5.4 Case 4: Histopathological classification of lung carcinoma 8
5.5 Case 5: Human cell line activation test (h-CLAT)-2 8
5.6 Case 6: Statistical model for predicting chemical toxicity 8

6 Statistical analysis for evaluating the precision of binary measurement methods and their results **9**
6.1 ISO 5725-based method 9
6.1.1 Overview 9
6.1.2 Case 1 12
6.1.3 Case 2(a) 12
6.1.4 Case 2(b) 13
6.1.5 Case 3(a) 13
6.1.6 Case 3(b) 13
6.2 Accordance and concordance 13
6.2.1 Overview 13
6.2.2 Case 1 15
6.2.3 Case 2(a) 15
6.2.4 Case 2(b) 16
6.2.5 Case 3(a) 16
6.2.6 Case 3(b) 16
6.3 ORDANOVA 16
6.3.1 Overview 16
6.3.2 Case 1 18
6.3.3 Case 2(a) 18
6.3.4 Case 2(b) 19
6.3.5 Case 3(a) 19
6.3.6 Case 3(b) 19
6.4 CM-accuracy, sensitivity and specificity 19
6.4.1 Overview 19
6.4.2 Case 4 20
6.4.3 Case 5 20
6.4.4 Case 6 20
6.5 Kappa coefficient 21
6.5.1 Overview 21
6.5.2 Case 4 21
6.5.3 Case 5 21
6.5.4 Case 6 22

7 Remarks on the methods introduced in this document **22**
7.1 Comparison between the mathematical expressions of the precision estimates 22
7.2 Comparison between the numerical examples of the precision estimates 23

Bibliography **25**

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 69, *Application of statistical methods*, Subcommittee SC 6, *Measurement methods and results*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

The documents in the ISO 5725 series define the precision of quantitative measurement methods and their results, and assume that the errors follow normal distributions in their basic models. Also, they provide how to run experiments to evaluate precision measures, such as repeatability and reproducibility. Nowadays, there is also a demand for dealing with qualitative measurement methods and their results, which output binary data, categorical data, etc. However, the ISO 5725 series is not suitable mathematically for analyzing such data.

Several existing studies propose statistical methods for dealing with binary and/or categorical data, but no guidance documents are available so far. Hence, this document summaries various methods to evaluate the precision of binary measurement methods and their results, which are the most essential and frequently used methods for qualitative data.

Statistical analysis for evaluating the precision of binary measurement methods and their results

1 Scope

This document introduces five statistical methods for evaluating the precision of binary measurement methods and their results. The five methods can be divided into two types. Both types are based on measured values provided by each laboratory participating in a collaborative study. In the first type, each laboratory repeatedly measures a single sample. The samples measured by the laboratories are nominally identical. The second type is an extension of the first type, where there are several levels of samples.

For each statistical method, this document briefly summarizes its theory and explains how to estimate the proposed precision measures. Some real cases are illustrated to help the readers understand the evaluation procedures involved. For the first and second types of methods, five and three cases are presented, respectively.

Finally, this document compares the five statistical methods.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 3534-1, *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*

ISO 5725-1, *Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions*

3 Terms and definitions, and symbols

3.1 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 3534-1 and ISO 5725-1 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1.1
accordance
probability that two binary measured values be identical when they are taken from the same laboratory

Note 1 to entry: The concept corresponds to the definition of “repeatability” in ISO 5725 and was originally proposed by Reference [3].

3.1.2
concordance
probability that two binary measured values be identical when they are taken from different laboratories

Note 1 to entry: The concept corresponds to the definition of "reproducibility" in ISO 5725 and was originally proposed by Reference [3].

3.1.3
ORDANOVA
statistical method for evaluating the precision of ordinal-scale measurement methods and their results based on an ordinal dispersion measure

Note 1 to entry: The concept was originally proposed by Reference [4].

3.1.4
true positive
TP
correct measured value in positive results, that is, the case where both the measured and the correct results are positive

3.1.5
true negative
TN
correct measured value in negative results, that is, the case where both the measured and the correct results are negative

3.1.6
false positive
FP
incorrect measured value in positive results, that is, the case where the measured value is positive but the correct one is negative

3.1.7
false negative
FN
incorrect measured value in negative results, that is, the case where the measured value is negative but the correct one is positive

3.1.8
confusion matrix
2×2 matrix showing the numbers of true positives, true negatives, false positives and false negatives

3.1.9
CM-accuracy
statistic for indicating the capability of two-class classifications, defined by the percentage of correct measured values

Note 1 to entry: The term CM-accuracy is identical to the term accuracy in the machine learning field, and is not generally used. However, this document uses CM-accuracy instead of accuracy in the machine learning field to distinguish between the term accuracy in ISO 5725 and the term in the machine learning field. In this document, the prefix CM stands for confusion matrixes because CM-accuracy can be calculated based on those.

3.1.10
sensitivity
statistic for indicating the capability of two-class classifications, defined by the percentage of correct positive measured values

3.1.11
specificity
statistic for indicating the capability of two-class classifications, defined by the percentage of correct negative measured values

3.1.12
CM-precision
statistic for indicating the capability of two-class classifications, defined by the percentage of correct measured values in positive measured values

Note 1 to entry: The term CM-precision is identical to the term precision in the machine learning field, and is not generally used. However, this document uses CM-precision instead of precision in the machine learning field to distinguish between the term precision in ISO 5725 and the term in the machine learning field. In this document, the prefix CM stands for confusion matrixes because CM-precision can be calculated based on those.

3.1.13
F-measure
statistic for indicating the capability of two-class classifications, defined by the harmonic mean between sensitivity and CM-precision

3.1.14
kappa coefficient
statistic for indicating the capability of two-class classifications, defined by the ratio of CM-accuracy minus the possibility of the correctness occurring by chance to one minus the possibility of the correctness occurring by chance

Note 1 to entry: Note to entry 1: The kappa coefficient is an extended statistic of CM-accuracy, originally introduced by Reference [15], which takes into account the possibility of the correctness occurring by chance.

3.2 Symbols

L number of laboratories participating in a collaborative study

n number of repetitions in each laboratory participating in a collaborative study

i suffix describing a laboratory, and $i \in \{1, \ldots, L\}$

j suffix describing a repetition, and $j \in \{1, \ldots, n\}$

y_{ij} measured value of repetition j of laboratory i

x_i sum of the measured values y_{ij} of laboratory i for the case where $y_{ij} \in \{0,1\}$, that is, $x_i = \sum_{j=1}^{n} y_{ij} \ (y_{ij} \in \{0,1\})$

$\bar{y}_i$ arithmetic mean of y_{ij} of laboratory i, that is, $\bar{y}_i = (1/n)\sum_{j=1}^{n} y_{ij} \ (i \in \{1, \ldots, L\})$

$\bar{\bar{y}}$ overall arithmetic mean of y_{ij}, that is, $\bar{\bar{y}} = (1/(nL))\sum_{i=1}^{L}\sum_{j=1}^{n} y_{ij}$

n_i^p number of positive measured values of laboratory i

n^p sum of n_i^p with respect to $i \in \{1, \ldots, L\}$, that is, $n^p = \sum_{i=1}^{L} n_i^p$

c_{ij} (i, j)-element of a confusion matrix

m general mean (expectation) in the basic model of ISO 5725-2

B_i laboratory component of bias under repeatability conditions of laboratory i in the basic model of ISO 5725-2

e_{ij} random error occurring in every measurement under repeatability conditions in the basic model of ISO 5725-2

σ_{ri}^2 within-laboratory variance of laboratory i

σ_r^2 repeatability variance or within-laboratory variance

σ_L^2 between-laboratory variance

σ_R^2 reproducibility variance, that is, $\sigma_R^2 = \sigma_r^2 + \sigma_L^2$

$\sigma_{ri:W}^2$ within-laboratory variance of laboratory i in the ISO 5725-based method

$\sigma_{r:W}^2$ repeatability variance in the ISO 5725-based method

$\sigma_{L:W}^2$ between-laboratory variance in the ISO 5725-based method

$\sigma_{R:W}^2$ reproducibility variance in the ISO 5725-based method, that is, $\sigma_{R:W}^2 = \sigma_{r:W}^2 + \sigma_{L:W}^2$

σ^2 ordinal dispersion measure proposed by Reference [14] for the case of binary data assumed to follow a binomial distribution

$\sigma_{ri:O}^2$ within-laboratory variance of laboratory i in ORDANOVA

$\sigma_{r:O}^2$ repeatability variance in ORDANOVA

$\sigma_{L:O}^2$ between-laboratory variance in ORDANOVA

$\sigma_{R:O}^2$ reproducibility variance in ORDANOVA, that is, $\sigma_{R:O}^2 = \sigma_{r:O}^2 + \sigma_{L:O}^2$

p_i probability of obtaining a measured value $y_{ij} = 1$ of laboratory i

p arithmetic mean of p_i, that is, $p = (1 / L)\sum_{i=1}^{L} p_i$

H_0 null hypothesis of a statistical test

χ_0^2 test statistic of a chi-squared test

A accordance of Reference [3] method

C concordance of Reference [3] method

$\hat{X}$ estimate of X

4 Overview

This document deals with the following five methods.

a) ISO 5725-based method (proposed by Reference [13]);

b) accordance and concordance (proposed by Reference [3]);

c) ORDANOVA (proposed by Reference [4]);

d) CM-accuracy, sensitivity and specificity;

e) Kappa coefficient (proposed by Reference [15]).

The assumed data structure depends on each method. In methods a), b) and c), each laboratory measures one identical sample multiple times, and the data can be summarized as in Table 1 and/or Table 2; while methods d) and e) are based on many levels of samples, and the data can be summarized as in Table 3.

NOTE Nowadays, the probability for detection (POD) approach is used to analyze binary measured values, instead of the methods d) and e), but this document deals with only classical methods. There are some ISO documents introducing the POD approach; see References [1] and [2].

Table 1 — Data format for methods a), b) and c)

Laboratory	Rep 1		Rep j		Rep n
Lab 1	y_{11}	⋯	y_{1j}	⋯	y_{1n}
⋮	⋮		⋮		⋮
Lab i	y_{i1}	⋯	y_{ij}	⋯	y_{in}
⋮	⋮		⋮		⋮
Lab L	y_{L1}	⋯	y_{Lj}	⋯	y_{Ln}
NOTE y_{ij} is either 0 or 1, which means negative and positive measured values, respectively.					

Table 2 — Another expression of Table 1

Laboratory	**Number of 1**	**Number of 0**
Lab 1	x_1	$n - x_1$
⋮	⋮	
Lab i	x_i	$n - x_i$
⋮	⋮	
Lab L	x_L	$n - x_L$
NOTE n is the number of repetitions in each laboratory, and $x_i \in \{0, 1, \ldots, n\}$.		

Table 3 — Data format for methods d) and e)

Actual values	**Measured values**	
	1	**0**
1	c_{11}	c_{12}
0	c_{21}	c_{22}
NOTE $c_{11}, c_{12}, c_{21}, c_{22}$ are non-negative integers.		

5 Examples used in this document

5.1 Case 1: Listeria monocytogenes

This subclause deals with the results of a collaborative study on Listeria monocytogenes, which was quoted and analyzed in Reference [13]. The study consisted of ten laboratories, where each laboratory repeated five-time measurements, that is, $L = 10,\ n = 5$. The results are shown in Table 4. In this table, numbers 1 and 0 mean that Listeria monocytogenes were and were not detected, respectively; and the column Total indicates the total number of detections of Listeria monocytogenes.

Table 4 — Case 1 — The results of a collaborative study on Listeria monocytogenes

Laboratory	Measured value					
	Repetition					Total
Lab 1	1	1	1	1	1	5
Lab 2	1	1	1	1	1	5
Lab 3	1	1	1	1	1	5
Lab 4	1	1	1	1	1	5
Lab 5	0	0	1	1	1	3
Lab 6	1	1	1	1	1	5
Lab 7	0	0	1	1	1	3
Lab 8	1	1	1	1	1	5
Lab 9	1	1	1	1	1	5
Lab 10	1	1	1	1	1	5

5.2 Case 2: Human cell line activation test (h-CLAT)-1

This subclause deals with the human cell line activation test (h-CLAT)[5][6]. The h-CLAT is an alternative to an animal experiment for evaluating the skin sensitization potential.

Reference [7] conducted a collaborative study and reported the results. The study consisted of five laboratories, where each laboratory repeated three-time measurements. Each laboratory measured 21 chemicals, but this document deals with two chemicals out of the 21, denoted by chemical *A* and chemical *B*, which have different pattern results. Case 2(a) reports the results of chemical *A*, shown in Table 5, and Case 2(b) reports the results of chemical *B*, shown in Table 6.

Table 5 — Case 2(a) — Number of detections of the skin sensitization potential of chemical *A* by h-CLAT in three-time measurements

Laboratory	Number of detections in three repetitions
Lab 1	3
Lab 2	3
Lab 3	1
Lab 4	3
Lab 5	3

Table 6 — Case 2(b) — Number of detections of the skin sensitization potential of chemical *B* by h-CLAT in three-time measurements

Laboratory	Number of detections in three repetitions
Lab 1	0
Lab 2	2
Lab 3	0
Lab 4	1
Lab 5	0

5.3 Case 3: Intraracheal administration testing

This subclause deals with the intratracheal administration test[8]. The test is a relatively new *in vivo* screening method for evaluating the pulmonary toxicity of nanomaterials.

The National Institute of Advanced Industrial Science and Technology (AIST) and five test laboratories conducted a collaborative study and reported the results[9]. The study consisted of five laboratories, where each laboratory reported positives (1) or negatives (0) of effects for 19 pathological findings at three doses using five rats. This document deals with two pathological findings at one dose, which have different patterns. Case 3(a) reports the results of appearance of alveolar macrophages following the administration of 0,13 mg/kg weight of a multi-wall carbon nanotube (MWCNT), shown in Table 7. Case 3(b) reports the results of hyperplasia of type II pneumocystis following the administration of 0,13 mg/kg weight of a MWCNT, shown in Table 8.

NOTE 1 See Reference [9] for the report, in Japanese, on the collaborative study and its original raw data.

NOTE 2 Each laboratory originally reported the strength of effects with five-level scores, −, ±, +, ++ and +++. In this document, these five-level scores are gathered to two-level scores, 1 and 0. When the original score was −, it is treated as negative (0); otherwise, the score is treated as positive (1).

Table 7 — Case 3 (a) —Results of a collaborative study of an intratracheal administration testing with appearance of alveolar macrophages

Laboratory	Number of rats reported in each category	
	0	**1**
Lab A	0	5
Lab B	0	5
Lab C	0	5
Lab D	0	5
Lab E	0	5

Table 8 — Case 3 (b) —Results of a collaborative study of an intratracheal administration testing with hyperplasia of type II pneumocystis

Laboratory	Number of rats reported in each category	
	0	**1**
Lab A	0	5
Lab B	3	2
Lab C	3	2
Lab D	1	4
Lab E	3	2

5.4 Case 4: Histopathological classification of lung carcinoma

Reference [10] compared 75-case diagnosis results of adenosquamous carcinoma, a type of lung carcinoma, rendered by three pathologists. The comparison results between two pathologists out of the three are shown in Table 9 as a confusion matrix. In this table, indices 0 and 1 mean grade II and grade III, respectively, and the non-negative integer in each cell stands for the number of cases.

Table 9 — Case 4 — Comparison results between two pathologists

	Pathologist 2	
Pathologist 1	**1**	**0**
1	27	4
0	3	41

5.5 Case 5: Human cell line activation test (h-CLAT)-2

Reference [11] compared the detection results of the h-CLAT and another alternative for evaluating the skin sensitization potential, the local lymph node assay (LLNA), using 117 chemicals. The comparison results are shown in Table 10 as a confusion matrix. In this table, indices 1 and 0 mean that a chemical was evaluated to have and not to have, respectively, the skin sensitization potential by each assay. The non-negative integer in each cell stands for the number of chemicals.

Table 10 — Case 5 — Comparison results between LLNA and h-CLAT

	h-CLAT	
LLNA	**1**	**0**
1	75	10
0	8	24

5.6 Case 6: Statistical model for predicting chemical toxicity

Some methods can be used for quantifying the prediction accuracy of statistical and machine learning models. Reference [12] developed a statistical model for predicting increased serum ALT levels in rats, which was one of the widely-used hepatotoxicity markers. The comparison results between the observed toxicity in rats and the predicted toxicity by the model are shown in Table 11 as a confusion matrix. In this table, indices 1 and 0 mean that a chemical was evaluated to have and not to have the hepatotoxicity, respectively. The non-negative integer in each cell stands for the number of chemicals.

Table 11 — Case 6 — Comparison results between observed and predicted results on increased serum ALT levels in rats

	Predicted values	
Observed values	**1**	**0**
1	18	5
0	39	114

6 Statistical analysis for evaluating the precision of binary measurement methods and their results

6.1 ISO 5725-based method

6.1.1 Overview

This method was originally introduced by Reference [13]. The study treated positive and negative measured values as integers one and zero, respectively, and directly applied the ISO 5725-2 approach to binary measured values.

The basic model of ISO 5725-2 is

$$y_{ij} = m + B_i + e_{ij}, \tag{1}$$

where y_{ij}, m and B_i are, respectively, the measured value of repetition $j \in \{1, \ldots, n\}$ in laboratory $i \in \{1, \ldots, L\}$, the general mean, and the laboratory component of variation in laboratory *i*. For any $i \in \{1, \ldots, L\}$, it assumes that the expectation of B_i is zero, $E(B_i) = 0$, and the variations of B_i are identical among all laboratories, $V(B_i) = \sigma_{ri}^2 = \sigma_r^2$. To estimate statistically the repeatability, between-laboratory and reproducibility variances, a one-way analysis of variance (one-way ANOVA) (random effects model) is performed. From Table 12, these variances are estimated as follows:

$$\hat{\sigma}_r^2 = s_I^2, \tag{2}$$

$$\hat{\sigma}_L^2 = \frac{s_{II}^2 - s_I^2}{n}, \tag{3}$$

and

$$\hat{\sigma}_R^2 = \hat{\sigma}_r^2 + \hat{\sigma}_L^2. \tag{4}$$

Table 12 — ANOVA table for ISO 5725-2

Source	Sum of squares (SQ)	Degree of freedom (df)	Mean square (MS) (=SQ/df)	Expected MS, *E*(*MS*)
Between lab.	$n\sum_{i=1}^{L}\left(\bar{y}_i - \bar{\bar{y}}\right)^2$	$L-1$	s_{II}^2	$n\sigma_L^2 + \sigma_r^2$
Within lab.	$\sum_{i=1}^{L}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2$	$L(n-1)$	s_I^2	σ_r^2
Total	$\sum_{i=1}^{L}\sum_{j=1}^{n}\left(y_{ij} - y\right)^2$	$Ln-1$		
$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$ $\left(i \in \{1, \ldots, L\}\right)$, that is, the arithmetic mean of y_{ij} of laboratory *i*; $\bar{\bar{y}} = \frac{1}{nL}\sum_{i=1}^{L}\sum_{j=1}^{n} y_{ij}$ $\left(= \frac{1}{L}\sum_{i=1}^{L}\bar{y}_i\right)$, that is, the overall arithemetic mean of y_{ij}.				

Reference [13] basic model is

$$y_{ij} = p + (p_i - p) + e_{ij}, \tag{5}$$

where y_{ij} is the measured value described as 0 (negative) or 1 (positive) for repetition $j \in \{1, \ldots, n\}$ in laboratory $i \in \{1, \ldots, L\}$, p_i and p are the probability of obtaining a measured value $y_{ij} = 1$ in laboratory i and its expectation, respectively.

Under the independence assumption of ISO 5725-2, y_{ij} $(j = 1, \cdots, n)$ in laboratory i follows a Bernoulli distribution with parameter p_i; thus, the within-laboratory variance of laboratory i is

$$\sigma_{ri:W}^2 = p_i (1 - p_i), \tag{6}$$

and the (whole) repeatability variance is defined as the average of σ_{ri}^2,

$$\sigma_{r:W}^2 = \frac{1}{L}\sum_{i=1}^{L}\sigma_{ri:W}^2 = \frac{1}{L}\sum_{i=1}^{L}p_i (1 - p_i) = p - \frac{1}{L}\sum_{i=1}^{L}p_i^2. \tag{7}$$

The between-laboratory variance in the population of L laboratories is defined as a classical variance of p_i, that is,

$$\sigma_{L:W}^2 = \frac{1}{L-1}\sum_{i=1}^{L}(p_i - p)^2, \tag{8}$$

and then the reproducibility variance is defined as the same as ISO 5725. In other words,

$$\sigma_{R:W}^2 = \sigma_{r:W}^2 + \sigma_{L:W}^2. \tag{9}$$

NOTE 1 Let $x_i = \sum_{j=1}^{n} y_{ij}$ be the number of positive measured values, then x_i follows a binomial distribution with parameters n and p_i.

From the definition of p_i and p, their estimates are calculated as

$$\hat{p}_i = \frac{1}{n}\sum_{j=1}^{n}y_{ij} \text{ and } \hat{p} = \frac{1}{L}\sum_{i=1}^{L}\hat{p}_i = \frac{1}{nL}\sum_{i=1}^{L}\sum_{j=1}^{n}y_{ij}, \tag{10}$$

respectively. In a similar way to ISO 5725-2, a modified one-way ANOVA is performed to estimate statistically the repeatability, between-laboratory and reproducibility variances. From Table 13, these variances are estimated as follows:

$$\hat{\sigma}_{r:W}^2 = s_I^2 = \frac{n}{L(n-1)}\sum_{i=1}^{L}\hat{p}_i (1 - \hat{p}_i), \tag{11}$$

$$\hat{\sigma}_{L:W}^2 = \frac{s_{II}^2 - s_I^2}{n} = \frac{1}{L-1}\sum_{i=1}^{L}(\hat{p}_i - \hat{p})^2 - \frac{1}{L(n-1)}\sum_{i=1}^{L}\hat{p}_i (1 - \hat{p}_i)\left(= \frac{1}{L-1}\sum_{i=1}^{L}\left(\widehat{\hat{p}_i} - \hat{p}\right)^2 - \frac{1}{n}\hat{\sigma}_{r:W}^2 \right), \tag{12}$$

and

$$\hat{\sigma}_{R:W}^2 = \hat{\sigma}_{r:W}^2 + \hat{\sigma}_{L:W}^2 . \tag{13}$$

Table 13 — ANOVA table for the Reference [13] method

Source	Sum of squares (SQ)	Degree of freedom (df)	Mean square (MS)(=SQ/df)	Expected MS, $E(MS)$
Between labs.	$n\sum_{i=1}^{L}\left(\hat{p}_i - \hat{p}\right)^2$	$L-1$	s_{II}^2	$n\sigma_{L:W}^2 + \sigma_{r:W}^2$
Within labs.	$n\sum_{i=1}^{L}\hat{p}_i\left(1-\hat{p}_i\right)$	$L(n-1)$	s_I^2	$\sigma_{r:W}^2$
Total	$Ln\,\hat{p}\left(1-\hat{p}\right)$	$nL-1$		

For conducting statistical tests to check whether the results of a collaborative study indicate different sensitivities p_i, Reference [13] proposed to apply the chi-squared test for independence in the contingency table shown in Table 14.

The null hypothesis is

$$H_0 : p_1 = p_2 = \cdots = p_L = p, \tag{14}$$

and the alternative hypothesis is that not all of the p_i are equal. Under the null hypothesis H_0 and the condition both $n\hat{p} \geq 5$ and $n\left(1-\hat{p}\right) \geq 5$ are satisfied, the following test statistic χ^2 is approximately chi-squared distributed with $L-1$ degrees of freedom.

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{L}\frac{\left(x_i - n\hat{p}\right)^2}{n\hat{p}} + \sum_{i=1}^{L}\frac{\left(\left(n-x_i\right) - n\left(1-\hat{p}\right)\right)^2}{n\left(1-\hat{p}\right)} = \sum_{i=1}^{L}\frac{n\left(\hat{p}_i - \hat{p}\right)^2}{\hat{p}} + \sum_{i=1}^{L}\frac{n\left(\hat{p}_i - \hat{p}\right)^2}{1-\hat{p}} \\ &= \frac{n}{\hat{p}\left(1-\hat{p}\right)}\sum_{i=1}^{L}\left(\hat{p}_i - \hat{p}\right)^2 .\end{aligned} \tag{15}$$

To check whether the null hypothesis H_0 is statistically rejected or not, the chi-squared test can be used. When the condition $n\hat{p} \geq 5$ and $n\left(1-\hat{p}\right) \geq 5$ is not satisfied, the Fisher's exact test can be used instead.

Table 14 — Contingency table for detecting a between-laboratory variance in the ISO 5725-based method

Attribute	Laboratory				Total
Number of positive measured values	x_1	x_2	$\cdots$	x_L	$\sum_{i=1}^{L}x_i$
Number of negative measured values	$n-x_1$	$n-x_2$	$\cdots$	$n-x_L$	$nL-\sum_{i=1}^{L}x_i$
Total	n	n		n	nL

NOTE 2 The number of repetitions $n \geq 10$ is necessary to satisfy the condition $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$; therefore, if $n < 10$ then the condition is never satisfied.

NOTE 3 Both the Fisher's exact test and the chi-squared test can be conducted using widely used statistical software such as R. For example, when one uses R, the former test can be conducted by a pre-install function, fisher.test(); while the latter test can be done by a pre-install function, chisq.test().

6.1.2 Case 1

The estimated detection probability $\hat{p}_i$ of each laboratory is listed in Table 15; therefore, $\hat{p} = 0,92$. Since $L = 10$ and $n = 5$, the estimates of the repeatability, between-laboratory and reproducibility variances are, respectively, calculated as follows:

$$\hat{\sigma}^2_{r:W} = \frac{5}{10 \cdot (5-1)} \left\{ \begin{array}{l} 1,0 \cdot (1-1,0) + 1,0 \cdot (1-1,0) + 1,0 \cdot (1-1,0) + 1,0 \cdot (1-1,0) + \\ 0,6 \cdot (0,6-1) + 1,0 \cdot (1-1,0) + 0,6 \cdot (1-0,6) + 1,0 \cdot (1-1,0) \\ +1, \cdot (1-1,0) + 1,0 \cdot (1-1,0) \end{array} \right\} = 0,060, \quad (16)$$

$$\hat{\sigma}^2_{L:W} = \frac{1}{10-1} \left\{ \begin{array}{l} (1,0-0,92)^2 + (1,0-0,92)^2 + (1,0-0,92)^2 \\ +(1,0-0,92)^2 + (0,6-0,92)^2 + (1,0-0,92)^2 \\ +(0,6-0,92)^2 + (1,0-0,92)^2 \\ +(1,0-0,92)^2 + (1,0-0,92)^2 \end{array} \right\} - \frac{1}{10 \cdot (5-1)} \cdot 0,060 \approx 0,016, \quad (17)$$

and

$$\hat{\sigma}^2_{R:W} = 0,060 + 0,016\ 4 \approx 0,076. \quad (18)$$

Because the *P*-value of the Fisher's exact test is 0,04, $H_0 : p_1 = p_2 = \cdots = p_L$ is rejected with 5 %-significance level; thus, a between laboratory variance is present from the viewpoint of statistics.

Table 15 — List of the estimated detection probability of each laboratory

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7	Lab 8	Lab 9	Lab 10
Estimated detection probability, $\hat{p}_i$	1,0	1,0	1,0	1,0	0,60	1,0	0,60	1,0	1,0	1,0

6.1.3 Case 2(a)

The estimated detection probability $\hat{p}_i$ of each laboratory is listed in Table 16; then, $\hat{p} = 0,866 \cong 0,87$. Since $L = 5$ and $n = 3$, the estimates of the repeatability, between-laboratory and reproducibility variances are, respectively, calculated as $\hat{\sigma}^2_{r:W} \approx 0,067$, $\hat{\sigma}^2_{L:W} \approx 0,067$ and $\hat{\sigma}^2_{R:W} \approx 0,13$.

Because the *P*-value of the Fisher's exact test is 0,14, $H_0 : p_1 = p_2 = \cdots = p_L$ is not rejected with 5 %-significance level.

Table 16 — List of the estimated detection probability of each laboratory

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
Estimated detection probability, $\hat{p}_i$	1,0	1,0	0,33	1,0	0,60

6.1.4 Case 2(b)

The estimated detection probability $\hat{p}_i$ of each laboratory is listed in Table 17; therefore, $\hat{p} = 0,20$. Since $L = 5$ and $n = 3$, the estimates of the repeatability, between-laboratory and reproducibility variances are, respectively, calculated as $\hat{\sigma}^2_{r:W} \approx 0,13$, $\hat{\sigma}^2_{L:W} \approx 0,044$ and $\hat{\sigma}^2_{R:W} \approx 0,18$.

Because the *P*-value of the Fisher's exact test is 0,41, $H_0 : p_1 = p_2 = \cdots = p_L$ is not rejected with 5 %-significance level.

Table 17 — List of the estimated detection probability of each laboratory

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
Estimated detection probability, $\hat{p}_i$	0,00	0,67	0,00	0,33	0,00

6.1.5 Case 3(a)

The estimated probability $\hat{p}_i$ of each laboratory is listed in Table 18; therefore, $\hat{p} = 1,0$. Since $L = 5$ and $n = 5$, the estimates of the repeatability, between-laboratory and reproducibility variances are, respectively, calculated as $\hat{\sigma}^2_{r:W} = 0,00$, $\hat{\sigma}^2_{L:W} = 0,00$ and $\hat{\sigma}^2_{R:W} = 0,00$.

Because the *P*-value of the Fisher's exact test is 1,0, $H_0 : p_1 = p_2 = \cdots = p_L$ is not rejected with 5 %-significance level.

Table 18 — List of the estimated detection probability of each laboratory

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
Estimated detection probability, $\hat{p}_i$	1,0	1,0	1,0	1,0	1,0

6.1.6 Case 3(b)

The estimated detection probability $\hat{p}_i$ of each laboratory is listed in Table 19; therefore, $\hat{p} = 0,60$. Since $L = 5$ and $n = 5$, the estimates of the repeatability, between-laboratory and reproducibility variances are, respectively, calculated as $\hat{\sigma}^2_{r:W} = 0,22$, $\hat{\sigma}^2_{L:W} = 0,036$ and $\hat{\sigma}^2_{R:W} \approx 0,26$.

Because the *P*-value of the Fisher's exact test is 0,19, $H_0 : p_1 = p_2 = \cdots = p_L$ is not rejected with 5 %-significance level.

Table 19 — List of the estimated detection probability of each laboratory

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
Estimated detection probability, $\hat{p}_i$	1,0	0,40	0,40	0,80	0,40

6.2 Accordance and concordance

6.2.1 Overview

Accordance and concordance were originally introduced by Reference [3]. They correspond to repeatability and reproducibility, respectively, in ISO 5725. These concepts were based on the probability that two measured values were identical.

The definition of accordance is the probability that pairs in each laboratory be identical; while that of concordance is the probability that pairs between different laboratories be identical. The estimates of accordance and concordance are, respectively, as follows:

$$\hat{A}_i \quad \left(\hat{A} \text{ of laboratory } i\right) = \frac{n_i^p\left(n_i^p - 1\right) + \left(n - n_i^p\right)\left(n - n_i^p - 1\right)}{n\left(n-1\right)}, \tag{19}$$

$$\hat{A} = \text{the arithmetric mean value of } \hat{A}_i, \tag{20}$$

and

$$\hat{C} = \frac{2n^p\left(n^p - nL\right) + nL\left(nL-1\right) - \hat{A}\cdot nL\left(n-1\right)}{n^2L\left(L-1\right)}, \tag{21}$$

where n, L and n_i^p are the number of the repetitions in each laboratory, that of laboratories, and that of positive measured values in laboratory i, respectively; and $n^p = \sum_{i=1}^{L} n_i^p$.

NOTE 1 The estimate of accordance of laboratory i is from the expression

$$\left(\binom{n_i^p}{2} + \binom{L - n_i^p}{2}\right) \Big/ \binom{n}{2}; \tag{22}$$

and that of concordance is from the expression

$$\frac{\binom{n^p}{2} + \binom{L - n^p}{2} - \hat{A}\cdot nL\left(n-1\right)}{n^2\binom{L}{2}}. \tag{23}$$

Each term in Formula (23) based on the number of pairs satisfying some conditions when arbitrary pairs of each laboratory measured values or all laboratory measured values are considered. The details are as follows:

$\binom{n^p}{2}$ is the number of pairs from all positive measured values in all laboratories n^p;

$\binom{L - n^p}{2}$ is the number of pairs from all negative measured values in all laboratories $L - n^p$;

$\hat{A}\cdot nL\left(n-1\right)$ is the sum of the number of pairs from positive measured values in each laboratory and pairs from negative measured values in each laboratory;

$n^2\binom{L}{2}$ is the number of pairs from all measured values in in all laboratories.

When concordance is less than accordance, a between-laboratory variance seems to be present; however, it is difficult to quantify the size of the between-laboratory variance. To demonstrate whether

a between-laboratory variance was present or not, Reference [3] proposed to consider concordance odds ratios (CORs), which were defined as follows:

$$\mathrm{COR} = \frac{100\,\hat{A}\cdot\left(100-100\,\hat{C}\right)}{100\,\hat{C}\cdot\left(100-100\,\hat{A}\right)}. \tag{24}$$

The COR is the odds ratio when the contingency table shown in Table 20 is considered; therefore, if $\mathrm{COR}=1$, then accordance = concordance, while if $\mathrm{COR}>1$, then accordance > concordance. To check whether $\mathrm{COR}=1$ or $\mathrm{COR}>1$, the Fisher's exact test or the chi-squared test can be used.

Table 20 — Contingency table for detecting a between-laboratory variance in Langton's method

Attribute	Number of pairs		Total
	of the same elements	of different elements	
Within laboratory	$100\,\hat{A}$	$100-100\,\hat{A}$	100
Between laboratory	$100\,\hat{C}$	$100-100\,\hat{C}$	100
Total	$100\,\left(\hat{A}+\hat{C}\right)$	$200-100\,\left(\hat{A}+\hat{C}\right)$	200

NOTE 2 Since $\hat{A}$ and $\hat{C}$ strongly depend on sensitivity, the difference between $\hat{C}$ and $\hat{A}$ is not suitable for an expression of the size of the between-laboratory variance.

6.2.2 Case 1

The estimate of accordance of each laboratory is shown in Table 21; therefore,

$$\hat{A} = \frac{1,0+1,0+1,0+1,0+0,40+1,0+0,40+1,0+1,0+1,0}{10} = 0,88. \tag{25}$$

Since the number of positive measured values in all laboratories n^p is 46, the estimate of concordance is calculated as follows:

$$\hat{C} = \frac{2\cdot 46\cdot\left(46-5\cdot 10\right)+5\cdot 10\cdot\left(5\cdot 10-1\right)-0,88\cdot 5\cdot 10\cdot\left(5-1\right)}{5^2\cdot 10\cdot\left(10-1\right)} \approx 0,85. \tag{26}$$

Then,

$$\mathrm{COR}\approx 1,3. \tag{27}$$

Because the *P*-value of the Fisher's exact test is 0,34, $H_0:\ \mathrm{COR}=1$ is not rejected with 5 %-significance level.

Table 21 — List of the estimate of accordance of each laboratory in Case 1

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7	Lab 8	Lab 9	Lab 10
$\hat{A}_i$	1,0	1,0	1,0	1,0	0,40	1,0	0,40	1,0	1,0	1,0

6.2.3 Case 2(a)

The estimate of accordance of each laboratory is shown in Table 22; therefore, $\hat{A}\approx 0,87$. Since $n^p=10$, $\hat{C}\approx 0,73$ and $\mathrm{COR}\approx 2,4$.

Because the *P*-value of the Fisher's exact test is 0,01, $H_0:\ \mathrm{COR}=1$ is rejected with 5 %-significance level; thus, a between-laboratory variance is present from the viewpoint of statistics.

Table 22 — List of the estimate of accordance of each laboratory in Case 2(a)

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
$\hat{A}_i$	1,0	1,0	0,33	1,0	1,0

6.2.4 Case 2(b)

The estimate of accordance of each laboratory is shown in Table 23; therefore, $\hat{A} \approx 0,73$. Since $n^p = 3$, $\hat{C} \approx 0,64$ and COR $\approx 1,5$.

Because the *P*-value of the Fisher's exact test is 0,11, $H_0: \text{COR} = 1$ is not rejected with 5 %-significance level.

Table 23 — List of the estimate of accordance of each laboratory in Case 2(b)

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
$\hat{A}_i$	0,00	0,67	0,00	0,33	0,00

6.2.5 Case 3(a)

The estimate of accordance of each laboratory is shown in Table 24; therefore, $\hat{A} = 1,0$. Since $n^p = 25$, $\hat{C} = 1,0$. In this case, the COR cannot be calculated since its denominator becomes 0. It is natural to consider that any between-laboratory variations are not present, because accordances of all laboratories are identical.

Table 24 — List of the estimate of accordance of each laboratory in Case 3(a

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
$\hat{A}_i$	1,0	1,0	1,0	1,0	1,0

6.2.6 Case 3(b)

The estimate of accordance of each laboratory is shown in Table 25; therefore, $\hat{A} = 0,56$. Since $n^p = 15$, $\hat{C} \approx 0,49$ and COR $\approx 1,3$.

Because the *P*-value of the Fisher's exact test is 0,20, $H_0: \text{COR} = 1$ is not rejected with 5 %-significance level.

Table 25 — List of the estimate of accordance of each laboratory in Case 3(b)

Laboratory	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
$\hat{A}_i$	1,0	0,40	0,40	0,60	0,40

6.3 ORDANOVA

6.3.1 Overview

ORDANOVA was originally introduced by Reference [4] for general ordinal-scale measurement methods and results; but this clause summarizes ORDANOVA only for binary cases.

ORDANOVA is based on an ordinal dispersion measure proposed by Reference [14], which is one of the most successful measure for an ordinal data variation. For binary data assumed to follow a binomial distribution with parameters n and q, the measure is defined as

$$\sigma^2 = 4q(1-q). \tag{28}$$

Using the measure, the total variance and the within-laboratory variance of laboratory i are, respectively, defined as

$$\sigma_T^2 = 4p(1-p) \text{ and } \sigma_{ri:O}^2 = 4p_i(1-p_i), \tag{29}$$

where p_i and p are, respectively, the probability of obtaining a measured value $y_{ij} = 1$ in laboratory i and its expectation. Also, using a classical variance of p_i, the between-laboratory variance in the population of L laboratories is defined as

$$\sigma_{L:O}^2 = \frac{4}{L}\sum_{i=1}^{L}(p_i - p)^2. \tag{30}$$

Because the following relationship holds among these variances:

$$\sigma_T^2 = \frac{1}{L}\sum_{i=1}^{L}\sigma_{ri:O}^2 + \sigma_{L:O}^2, \tag{31}$$

the repeatability and reproducibility variances are, respectively, defined as the average of $\sigma_{ri:O}^2$ and the total variance. In other words,

$$\sigma_{r:O}^2 = \frac{1}{L}\sum_{i=1}^{L}\sigma_{ri:O}^2 \text{ and } \sigma_{R:O}^2 = \sigma_T^2. \tag{32}$$

From the definitions of p_i and p, their estimates are, respectively, calculated as

$$\hat{p}_i = \frac{n_i^p}{n} \text{ and } \hat{p} = \frac{1}{L}\sum_{i=1}^{L}\hat{p}_i = \frac{1}{nL}\sum_{i=1}^{L}n_i^p, \tag{33}$$

where n_i^p stands for the number of positive measured values in laboratory i. Then, the repeatability, between-laboratory and reproducibility variances are, respectively, estimated as

$$\hat{\sigma}_{r:O}^2 = \frac{1}{L}\sum_{i=1}^{L}4\hat{p}_i(1-\hat{p}_i) = \frac{4}{L}\sum_{i=1}^{L}\frac{n_i^p}{n}\left(1-\frac{n_i^p}{n}\right), \tag{34}$$

$$\hat{\sigma}_{L:O}^2 = \frac{4}{L}\sum_{i=1}^{L}(\hat{p}_i - \hat{p})^2 = \frac{4}{L}\sum_{i=1}^{L}\left(\frac{n_i^p}{n} - \frac{1}{nL}\sum_{i=1}^{L}n_i^p\right)^2, \tag{35}$$

and

$$\hat{\sigma}_{R:O}^2 = 4\hat{p}(1-\hat{p}) = 4\left(\frac{1}{L}\sum_{i=1}^{L}\hat{p}_i\right)\left(1-\frac{1}{L}\sum_{i=1}^{L}\hat{p}_i\right) = 4\left(\frac{1}{nL}\sum_{i=1}^{L}n_i^p\right)\left(1-\frac{1}{nL}\sum_{i=1}^{L}n_i^p\right). \tag{36}$$

For conducting statistical tests to check whether the results of a collaborative study indicate different sensitivities p_i, Reference [4] applied the same method as Reference [13]. More precisely, they introduced the statistic I and proved the following relationships:

$$I = \frac{\hat{\sigma}_L^2 / df_B}{\hat{\sigma}_R^2 / df_T} = \frac{nL-1}{nL}\frac{\chi^2}{L-1} \cong \frac{\chi^2}{L-1}. \tag{37}$$

Then, they proposed that the null hypothesis is rejected if

$$I > \frac{\chi_{\alpha-1}^2}{L-1}, \tag{38}$$

where $\chi_{\alpha-1}^2$ stands for the critical value of the chi-square statistic.

NOTE 1 The reproducibility variance can be estimated as follows, too:

$$\begin{aligned}\hat{\sigma}_{R:O}^2 &= \frac{1}{L}\sum_{i=1}^{L}\hat{\sigma}_{ri:O}^2 + \hat{\sigma}_{L:O}^2 = \frac{1}{L}\sum_{i=1}^{L}4\hat{p}_i\left(1-\hat{p}_i\right) + \frac{4}{L}\sum_{i=1}^{L}\left(\hat{p}_i-\hat{p}\right)^2 \\ &= \frac{4}{L}\sum_{i=1}^{L}\left\{\frac{n_i^p}{n}\left(1-\frac{n_i^p}{n}\right) + \left(\frac{n_i^p}{n} - \frac{1}{nL}\sum_{i=1}^{L}n_i^p\right)^2\right\}.\end{aligned} \tag{39}$$

NOTE 2 The statistical test for checking whether the null hypothesis is statistically rejected or not is completely identical to the method introduced by Reference [13]; therefore, when the condition $n\hat{p} \geq 5$ and $n\left(1-\hat{p}\right) \geq 5$ is not satisfied, the Fisher's exact test can be used instead.

6.3.2 Case 1

The estimated probability $\hat{p}_i$ of each laboratory is listed in Table 15; therefore, $\hat{p} = 0,92$. Since $L = 10$ and $n = 5$, the estimates of the within-laboratory, between-laboratory and reproducibility variances are, respectively, calculated as follows:

$$\hat{\sigma}_r^2 = \frac{4}{10}\left\{\begin{array}{l}1,0\cdot\left(1-1,0\right)+1,0\cdot\left(1-1,0\right)+1,0\cdot\left(1-1,0\right)+1,0\cdot\left(1-1,0\right)\\ +0,6\cdot\left(0,6-1\right)+1,0\cdot\left(1-1,0\right)+0,6\cdot\left(1-0,6\right)+1,0\cdot\left(1-1,0\right)\\ +1,0\cdot\left(1-1,0\right)+1,0\cdot\left(1-1,0\right)\end{array}\right\} \approx 0,19, \tag{40}$$

$$\hat{\sigma}_L^2 = \frac{1}{10}\left\{\begin{array}{l}\left(1,0-0,92\right)^2+\left(1,0-0,92\right)^2+\left(1,0-0,92\right)^2+\left(1,0-0,92\right)^2\\ +\left(0,60-0,92\right)^2+\left(1,0-0,92\right)^2+\left(0,60-0,92\right)^2\\ +\left(1,0-0,92\right)^2+\left(1,0-0,92\right)^2+\left(1,0-0,92\right)^2\end{array}\right\} \approx 0,10, \tag{41}$$

and

$$\hat{\sigma}_R^2 = 4\ \left(0,92\ \left(1-0,92\right)\right) \approx 0,29. \tag{42}$$

6.3.3 Case 2(a)

The estimated probability $\hat{p}_i$ of each laboratory is listed in Table 16; therefore, $\hat{p} \approx 0,87$. Since $L = 5$ and $n = 3$, the estimates of the within-laboratory, between-laboratory and reproducibility variances are, respectively, calculated as $\hat{\sigma}_r^2 \approx 0,18$, $\hat{\sigma}_L^2 \approx 0,28$ and $\hat{\sigma}_R^2 \approx 0,46$.

6.3.4 Case 2(b)

The estimated probability $\hat{p}_i$ of each laboratory is listed in Table 17; therefore, $\hat{p} = 0,20$. Since $L = 5$ and $n = 3$, the estimates of the within-laboratory, between-laboratory and reproducibility variances are , respectively, calculated as $\hat{\sigma}_r^2 \approx 0,36$, $\hat{\sigma}_L^2 \approx 0,28$ and $\hat{\sigma}_R^2 \approx 0,64$.

6.3.5 Case 3(a)

The estimated probability $\hat{p}_i$ of each laboratory is listed in Table 18; therefore, $\hat{p} = 1,0$. Since $L = 5$ and $n = 5$, the estimates of the within-laboratory, between-laboratory and reproducibility variances are, respectively, calculated as $\hat{\sigma}_r^2 = 0,00$, $\hat{\sigma}_L^2 = 0,00$ and $\hat{\sigma}_R^2 = 0,00$.

6.3.6 Case 3(b)

The estimated probability $\hat{p}_i$ of each laboratory is listed in Table 19; therefore, $\hat{p} = 0,60$. Since $L = 5$ and $n = 5$, the estimates of the within-laboratory, between-laboratory and reproducibility variances are, respectively, calculated as $\hat{\sigma}_r^2 \approx 0,70$, $\hat{\sigma}_L^2 \approx 0,26$ and $\hat{\sigma}_R^2 \approx 0,96$.

6.4 CM-accuracy, sensitivity and specificity

6.4.1 Overview

CM-accuracy, sensitivity and specificity are widely used statistics based on the true value of 1 or 0 (or positive/negative). These statistics can be used for quantifying the capacities of a laboratory, of an alternative measurement method, of a statistical model by comparing the true values, and so on.

These are defined as follows:

$$\text{CM-Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}(= \text{TP} + \text{TN} + \text{FP} + \text{FN})}, \tag{43}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{44}$$

and

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{45}$$

where TP, TN, FP and FN stand for true positive, true negative, false positive and false negative, respectively.

NOTE 1 Sensitivity represents the percentage of positive measured values to actual positives; while specificity represents the percentage of negative measured values to actual negatives. The two statistics represent the capability of measurement methods for determining 'True' or 'False'; see Table 26.

Table 26 — Confusion matrix

Actual	Measurement		
	True	False	Sum
True	True positive (TP)	False negative (FN)	TP+FN
False	False positive (FP)	True negative (TN)	FP+TN
Sum	TP+FP	FN+TN	Total

There are several similar kinds of statistics to CM-accuracy, sensitivity and specificity; for example, CM-precision, F-measure and balanced accuracy. These definitions are listed as follows:

$$\text{CM-precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{46}$$

$$\text{F-measure} = \frac{2\text{sensitivity} \cdot \text{CM-precision}}{\text{sensitivity} + \text{CM-precision}}, \tag{47}$$

and

$$\text{Balanced accuracy} = \frac{\text{sensitivity} + \textit{s}\text{pecificity}}{2}. \tag{48}$$

NOTE 2 CM-precision is defined as the percentage of actual positives to positive measured values. The statistic represents the completeness of the detection of positives.

NOTE 3 F-measure is the harmonic mean between sensitivity and CM-precision. Since sensitivity and CM-precision describe, respectively, the correctness and the completeness of the detection of positives, the F-measure is one of the integrated statistics of the detection of positives.

NOTE 4 Balanced accuracy is the arithmetic mean between sensitivity and specificity, so the statistic describes the accuracy not depending on the distribution of the true values.

6.4.2 Case 4

The results are as follows:

$$\text{CM-accuracy} = \frac{27+41}{75} \approx 0,91. \tag{49}$$

$$\text{Sensitivity} = \frac{27}{27+4} \approx 0,87. \tag{50}$$

$$\text{Specificity} = \frac{41}{41+8} \approx 0,93 \tag{51}$$

$$\text{CM-precision} = \frac{27}{27+3} = 0,90. \tag{52}$$

$$\text{F-measure} = \frac{2 \cdot 0,870 \cdot 0,900}{0,870+0,900} \approx 0,88. \tag{53}$$

$$\text{Balanced accuracy} = \frac{0,870+0,931}{2} \approx 0,90. \tag{54}$$

6.4.3 Case 5

The results are CM-accuracy ≈ 0,85, sensitivity ≈ 0,88, specificity = 0,75, CM-precision ≈ 0,90, F-measure ≈ 0,89 and balanced accuracy ≈ 0,82.

6.4.4 Case 6

The results are CM-accuracy = 0,75, sensitivity ≈ 0,78, specificity ≈ 0,75, CM-precision ≈ 0,32, F-measure = 0,45 and balanced accuracy ≈ 0,76.

6.5 Kappa coefficient

6.5.1 Overview

CM-accuracy has a problem that it strongly depends on the distribution of the true values. To overcome the problem, Reference [15] introduced the kappa coefficient.

The kappa coefficient is defined as

$$\kappa = \frac{P - P_e}{1 - P_e}, \tag{55}$$

where P and P_e stand, respectively, for CM-accuracy (overall agreement) and the probability of the agreement occurring by chance. P_e is predicted by the following calculation:

$$P_e = \frac{(\mathrm{TP} + \mathrm{FN})}{\mathrm{Total}} \frac{(\mathrm{TP} + \mathrm{FP})}{\mathrm{Total}} + \frac{(\mathrm{FP} + \mathrm{TN})}{\mathrm{Total}} \frac{(\mathrm{FN} + \mathrm{TN})}{\mathrm{Total}}. \tag{56}$$

Several criteria were proposed to determine the value of the kappa coefficient κ. Table 27 shows some of the criteria.

Table 27 — Proposed criteria for the kappa coefficient

κ	Landis et al.	Cicchetti et al.	Fleiss
≤ 0	poor	poor	poor
0,00 to 0,20	slight		
0,21 to 0,40	fair		
0,41 to 0,60	moderate	fair	fair to good
0,61 to 0,75	substantial	excellent	
0,75 to 0,80			excellent
0,81 to 1,0	almost perfect		

6.5.2 Case 4

CM-accuracy and the probability of the agreement occurring by chance are, respectively,

$$P = \frac{27+41}{75} = 0,906 \tag{57}$$

and

$$P_e = \frac{(27+4)}{75} \frac{(27+3)}{75} + \frac{(3+41)}{75} \frac{(4+41)}{75} = 0,517; \tag{58}$$

therefore, the kappa coefficient is

$$\kappa = \frac{0,906 - 0,517}{1 - 0,517} \approx 0,81. \tag{59}$$

6.5.3 Case 5

CM-accuracy and the probability of the agreement occurring by chance are $P = 0,846$ and $P_e = 0,595$, respectively; therefore, the kappa coefficient is $\kappa \approx 0,62$.

6.5.4 Case 6

CM-accuracy and the probability of the agreement occurring by chance are $P = 0,750$ and $P_e = 0,630$, respectively; therefore, the kappa coefficient is $\kappa \approx 0,32$.

7 Remarks on the methods introduced in this document

7.1 Comparison between the mathematical expressions of the precision estimates

This subclause compares the proposed mathematical expressions of the precision estimates introduced in this document.

First, assuming that the estimated values of the repeatability, between-laboratory and reproducibility variances in the ISO 5725-based method are given, then one can obtain those of ORDANOVA by using the following relationships:

$$\hat{\sigma}^2_{r:O} = 4\left(\frac{n-1}{n}\right)\hat{\sigma}^2_{r:w}, \tag{60}$$

$$\hat{\sigma}^2_{L:O} = 4\left(\frac{L-1}{L}\hat{\sigma}^2_{L:W} + \frac{L-1}{nL}\hat{\sigma}^2_{r:W}\right), \tag{61}$$

$$\hat{\sigma}^2_{R:O} = 4\left(\frac{L-1}{L}\hat{\sigma}^2_{R:W} + \frac{n-1}{nL}\hat{\sigma}^2_{r:W}\right); \tag{62}$$

and one can obtain the estimated values of accordance and concordance of Langton's method by using the following relationships:

$$\hat{A} = 1 - 2\hat{\sigma}^2_{r:W}, \tag{63}$$

$$\hat{C} = 1 - 2\hat{\sigma}^2_{R:W}. \tag{64}$$

Second, assuming that the estimated values of accordance and concordance of Langton's method are given, one can obtain the estimated values of the three precision statistics in the ISO-based method by using the following relationships:

$$\hat{\sigma}^2_{r:W} = \frac{1-\hat{A}}{2}, \tag{65}$$

$$\hat{\sigma}^2_{L:W} = \frac{\hat{A}-\hat{C}}{2}, \tag{66}$$

$$\hat{\sigma}^2_{R:W} = \frac{1-\hat{C}}{2}; \tag{67}$$

and one can obtain those of ORDANOVA by using the following relationships:

$$\hat{\sigma}^2_{r:O} = \frac{2(n-1)}{n}\left(1-\hat{A}\right), \tag{68}$$

$$\hat{\sigma}^2_{L:O} = \frac{2(L-1)}{L}\left(1-\hat{C}\right) - \frac{2(n-1)(L-1)}{nL}\left(1-\hat{A}\right), \tag{69}$$

$$\hat{\sigma}^2_{R:O} = \frac{2(L-1)}{L}(1-\hat{C}) - \frac{2(n-1)}{nL}(1-\hat{A}). \tag{70}$$

Third, assuming that the estimate values of the repeatability, between-laboratory and reproducibility variances in ORDANOVA are given, then one can obtain those of the ISO 5725-based method by using the following relationships:

$$\hat{\sigma}^2_{r:W} = \frac{1}{4}\left(\frac{n}{n-1}\right)\hat{\sigma}^2_{r:O}, \tag{71}$$

$$\hat{\sigma}^2_{L:W} = \frac{1}{4}\left\{\left(\frac{L}{L-1}\right)\hat{\sigma}^2_{L:O} - \left(\frac{1}{n-1}\right)\hat{\sigma}^2_{r:O}\right\}, \tag{72}$$

$$\hat{\sigma}^2_{R:W} = \frac{1}{4}\left\{\left(\frac{L}{L-1}\right)\hat{\sigma}^2_{R:O} + \left(\frac{1}{L-1}\right)\hat{\sigma}^2_{r:O}\right\}; \tag{73}$$

and one can obtain the estimated values of accordance and concordance of Langton's method by using the following relationships:

$$\hat{A} = 1 - \frac{1}{2}\left(\frac{n}{n-1}\right)\hat{\sigma}^2_{r:O}, \tag{74}$$

$$\hat{C} = 1 - \frac{1}{2}\left\{\left(\frac{L}{L-1}\right)\hat{\sigma}^2_{R:O} + \left(\frac{1}{L-1}\right)\hat{\sigma}^2_{r:O}\right\}. \tag{75}$$

These relationships show that the three methods, the ISO 5725-based method, Langton's method and ORDANOVA, are essentially identical. These three methods give a re-expression of the same results.

NOTE The comparison results between the estimates in the ISO 5725-based method and Langton's method were originally presented in Reference [16].

Some remarks on the statistics related to confusion matrices.

CM-accuracy, sensitivity and specificity are widely used in several fields. When the numbers of positive samples and negative samples are almost identical, these statistics can be useful to quantify the precision of binary measurement methods and their results. However, when there is a large difference in numbers between positive and negative samples, these statistics may not describe precision properly. Under the condition that the number of positive samples is 90 and that of negative samples is 10, if a binary measurement method always outputs the positive value, then CM-accuracy = 0,90, sensitivity =0,10 and specificity =1,0. These imply that the measurement method is sufficiently useful for determining positive samples, even though the measurement method determined nothing. In such cases, balanced accuracy and the kappa coefficient are useful. Since the former is the arithmetic mean between sensitivity and specificity, its meaning is easy to understand. Also, the latter is a mathematical extension of CM-accuracy, which takes into account the possibility of the correctness occurring by chance, and several criteria to understand values of the kappa coefficient were proposed.

In the case where the measurement precision on positive samples is important, CM-precision and F-measure would be useful in addition to sensitivity. CM-precision describes how large a measurement method covers for positive samples. F-measure is an integrated statistic for the detection of positive samples, since it is defined as the harmonic mean between sensitivity and precision.

7.2 Comparison between the numerical examples of the precision estimates

Table 28 to Table 32 summarize the results of the examples for each case.

Regarding the statistical tests for detecting between-laboratory variances, the ISO 5725-based method and ORDANOVA apply a chi-squared test for the completely identical contingency table, but Langton's method uses a different contingency table. Therefore, the results of the former two methods are

identical, but Langton's method has a possibility of showing a different conclusion from the other two methods.

Table 28 — Results of Case 1

Method	Statistics			
	Repeatability variances / Accordance	Between-laboratory variances / COR	Reproducibility variances / Concordance	Result of statistical test with 5 %-significance level
ISO 5725-based method	0,060	0,016	0,076	rejected
Langton's method	0,88	1,3	0,85	not rejected
ORDANOVA	0,19	0,10	0,29	rejected

Table 29 — Results of Case 2(a)

Method	Statistics			
	Repeatability variances / Accordance	Between-laboratory variances / COR	Reproducibility variances / Concordance	Result of statistical test with 5 %-significance level
ISO 5725-based method	0,067	0,067	0,13	not rejected
Langton's method	0,87	2,4	0,73	rejected
ORDANOVA	0,18	0,28	0,46	not rejected

Table 30 — Results of Case 2(b)

Method	Statistics			
	Repeatability variances / Accordance	Between-laboratory variances / COR	Reproducibility variances / Concordance	Result of statistical test with 5 %-significance level
ISO 5725-based method	0,13	0,044	0,18	not rejected
Langton's method	0,73	1,5	0,64	not rejected
ORDANOVA	0,36	0,28	0,64	not rejected

Table 31 — Results of Case 3(a)

Method	Statistics			
	Repeatability variances / Accordance	Between-laboratory variances / COR	Reproducibility variances / Concordance	Result of statistical test with 5 %-significance level
ISO 5725-based method	0,00	0,00	0,00	not rejected
Langton's method	1,0	-	1,0	-
ORDANOVA	0,00	0,00	0,00	not rejected

Table 32 — Results of Case 3(b)

Method	Statistics			
	Repeatability variances / Accordance	Between-laboratory variances / COR	Reproducibility variances / Concordance	Result of statistical test with 5 %-significance level
ISO 5725-based method	0,22	0,036	0,26	not rejected
Langton's method	0,56	1,3	0,49	not rejected
ORDANOVA	0,70	0,26	0,96	not rejected

Bibliography

[1] ISO 16140 (all parts), *Microbiology of the food chain - Method validation*

[2] ISO/TS 27878, *Reproducibility of the LOD of binary methods in collaborative and in-house validation studies*

[3] Langton S.D., Chevennement R., Nagelkerke N., Lombard B., Analysing collaborative trials for qualitative microbiological methods: accordance and concordance, Int. J. Food Microbiol. **79** (2002) 175–181. https://doi.org/10.1016/s0168-1605(02)00107-1.

[4] Gadrich T., Bashkansky E., ORDANOVA: Analysis of ordinal variation, J. Stat. Plan. Inference. **142** (2012) 3174–3188. https://doi.org/10.1016/j.jspi.2012.06.004.

[5] Ashikaga T., Yoshida Y., Hirota M., Yoneyama K., Itagaki H., Sakaguchi H., Miyazawa M., Ito Y., Suzuki H., Toyoda H., Development of an in vitro skin sensitization test using human cell lines: The human Cell Line Activation Test (h-CLAT), Toxicol. In Vitro. **20** (2006) 767–773. https://doi.org/10.1016/j.tiv.2005.10.012.

[6] Sakaguchi H., Ashikaga T., Miyazawa M., Yoshida Y., Ito Y., Yoneyama K., Hirota M., Itagaki H., Toyoda H., Suzuki H., Development of an in vitro skin sensitization test using human cell lines; human Cell Line Activation Test (h-CLAT) II. An inter-laboratory study of the h-CLAT, Toxicol. In Vitro. **20** (2006) 774–784. https://doi.org/10.1016/j.tiv.2005.10.014.

[7] Sakaguchi H., Ryan C., Ovigne J.-M., Schroeder K.R., Ashikaga T., Predicting skin sensitization potential and inter-laboratory reproducibility of a human Cell Line Activation Test (h-CLAT) in the European Cosmetics Association (COLIPA) ring trials, Toxicol. In Vitro. **24** (2010) 1810–1820. https://doi.org/10.1016/j.tiv.2010.05.012.

[8] Driscoll K.E., Costa D.L., Hatch G., Henderson R., Oberdorster G., Salem H., Schlesinger R.B., Intratracheal Instillation as an Exposure Technique for the Evaluation of Respiratory Tract Toxicity: Uses and Limitations, Toxicol. Sci. **55** (2000) 24–35. https://doi.org/10.1093/toxsci/55.1.24.

[9] AIST, Annual Report on the Project, "Survey on standardization of intratracheal administration study for nanomaterials and related issues" (2017), (2018). https://www.meti.go.jp/meti_lib/report/H29FY/000102.pdf (accessed March 19, 2020).

[10] Ghandur-Mnaymne L., Raub W.A., Sidhar K.S., orge Albores-Saavedr, E. Gould, R.C. Duncan, The accuracy of the histological classification of lung carcinoma and its reproducibility: A study of 75 archival cases of adenosquamous carcinoma, Cancer Invest. **11** (1993) 641–651. https://doi.org/10.3109/07357909309046936.

[11] Ashikaga T., Sakaguchi H., Sono S., Kosaka N., Ishikawa M., Nukada Y., Miyazawa M., Ito Y., Nishiyama N., Itagaki H., A Comparative Evaluation of In Vitro Skin Sensitisation Tests: The Human Cell-line Activation Test (h-CLAT) versus the Local Lymph Node Assay (LLNA), Altern. Lab. Anim. **38** (2010) 275–284. https://doi.org/10.1177/026119291003800403.

[12] Takeshita J., Nakayama H., Kitsunai Y., Tanabe M., Oki H., Sasaki T., Yoshinari K., Discriminative models using molecular descriptors for predicting increased serum ALT levels in repeated-dose toxicity studies of rats, Comput. Toxicol. **6** (2018) 64–70. https://doi.org/10.1016/j.comtox.2017.05.002.

[13] Wilrich P.-Th., The determination of precision of qualitative measurement methods by interlaboratory experiments, Accreditation Qual. Assur. **15** (2010) 439–444. https://doi.org/10.1007/s00769-010-0661-1.

[14] Blair J., Lacy M.G., Statistics of ordinal variation, Sociol. Mathods Res. **28** (2000) 251–280.

[15] COHEN J., A coefficient of agreement for nominal scales, Educ. Psychol. Meas. **20** (1960) 37–46. https://doi.org/10.1177/001316446002000104.

[16] SUZUKI T., TSUTSUMI Y., KAWAMURA H., Viewpoints to characterize precision evaluation methods in binary measurements, Measurement. **46** (2013) 3710–3714. https://doi.org/10.1016/j.measurement.2013.05.032.

ICS 03.120.30

Price based on 26 pages