TECHNICAL SPECIFICATION

ISO/TS 16393

First edition
2019-02

Molecular biomarker analysis — Determination of the performance characteristics of qualitative measurement methods and validation of methods

Analyse de biomarqueurs moléculaires — Détermination des caractéristiques de performance des méthodes de mesure qualitatives et validation des méthodes

Reference number
ISO/TS 16393:2019(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

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 34, *Food products*, Subcommittee SC 16, *Horizontal methods for molecular biomarker analysis*.

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

Qualitative (binary) analytical methods (e.g. applied to screening tests) for use in the analysis of food or food products (including seeds of food crops) with the purpose of demonstrating the presence/absence of a given measurand in a sample should provide objective evidence that they are adequate for their intended use. A validated test method is much preferred over one that has not undergone studies to determine its accuracy and reliability for its specific purpose. These methods that yield a binary result (yes/no, positive/negative, etc.) are referred to as "qualitative" or "binary" methods.

As with quantitative methods, qualitative method performance has to be characterized with respect to the concentration of the measurand. However, only two conditions are indicated in the result; either the measurand is detected (a positive result) or it is not detected (a negative result). While internationally recognized guidelines (e.g. ISO 5725-2, References [7] and [16]) have been produced over the years to harmonize the validation of quantitative analytical methods, no consensus is yet available among stakeholders on a practical implementation of the performance criteria approach to the validation of qualitative methods for use in food and food products.

Conceptual approaches for validating qualitative methods classically focused on parameters such as sensitivity, selectivity, false positive rate and false negative rate, based on detection/non-detection of the measurand in the test sample. The limitation of this approach was the underlying assumption that the method had a predictable response to the presence of a measurand present at a non-zero concentration. In practice, however, a non-zero concentration can result in a variable probability of a positive result in the assay. Treating the concentration of measurand as a continuous variable with reasonable and/or previously determined confidence in a defined matrix using a specific analytical method is a better predictor of measurement response than a two-state, zero/non-zero variable.

This document describes the assessment of probability of detection (POD). This approach allows for comparison of probabilities across concentrations and further allows for a simple graphical representation of validation data as a POD response curve graphed by concentration with associated error bars of the mean POD value. This approach expresses the POD as dependent on concentration; the goal of validation is to characterize the response probability curve as a function of measurand mass or concentration.

A number of models have been described in the literature for the calculations of the confidence intervals of the POD and confidence intervals or predictive ranges for concentrations in case of a positive or negative result, e.g. References [4], [8], [9], [11], [17], [19] and [20]. Whereas qualitative methods are often evaluated at 50 %, they are used close to 100 %, or at levels where the sample size is adjusted so as to always obtain a clear positive or negative result. The present specification is therefore the result of an extensive discussion of the possible improved models for characterization of qualitative methods, particularly focused on the characterization of the methods close to the 0 and 100 % POD cases. The performance characteristics include:

a) the mean POD across laboratories (LPOD);

b) the corresponding confidence interval of the LPOD, which is the interval estimate of the mean POD;

c) the prediction interval for future observations of laboratory specific PODs.

An advanced statistical method allows the user to calculate confidence and/or prediction intervals for the concentrations where the user would expect positive or negative results. To do so is particularly challenging where the POD is close to 0 % or 100 %.

Molecular biomarker analysis — Determination of the performance characteristics of qualitative measurement methods and validation of methods

1 Scope

This document specifies methods that yield a binary result and are used for the determination in food or food products (including seeds of food crops) of the presence of molecular biomarkers. These methods are typically used where the measurand is expected to be present in very small amounts and concentrations at the limit of detection (LOD).

Methods are validated in terms of the probability of detection (POD) and of the precision of the POD. They do not rely on the concept of false positive/false-negative results, or the concept of LOD. However, inferences about the precision of the classical LOD can be made.

This document describes the extent of method validation. The annexes provide different statistical models that can be considered depending on the analytical method, structure of data and statistical experience.

This document does not apply to quantitative methods that are used to make a detection decision by comparing the value of a response to a cut-off value using a quantitative method, where the methods are validated by using quantitative statistics on the responses. This document also does not apply to microbiological test methods, starch, essential oils or quantitative 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 5725-1:1994, *Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions*

ISO 5725-2:1994, *Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method*

3 Terms and definitions

For the purposes of this document, the following terms and definitions 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 http://www.electropedia.org/

3.1
binary result
result from a *method* (3.6) of analysis where there are only two possible outcomes

3.2
intraclass correlation coefficient
ICC
measure of the reliability of measurements (between laboratories)

Note 1 to entry: The coefficient represents agreements between two or more results measured on identical samples.

3.3
identical test item
sample that is prepared and can be presumed to be identical for the intended purpose of measurement of the measurand (and can be presumed to be identical for the intended purpose)

[SOURCE: ISO 3534-2:2006, 1.2.34, modified — "and can be presumed to be identical purpose of measurement of the measurand" has been added and Note 1 to entry has been deleted.]

3.4
lower confidence limit
LCL
$\hat{\mu}_L$
lower value of a range containing the true value of the measurand with a specified probability

Note 1 to entry: The symbol for LCL is taken from Reference [5].

3.5
mean probability of detection across laboratories
LPOD
$P_{\alpha\lambda}$
probability of a positive analytical outcome of a *qualitative method* (3.9) for a given matrix at a given concentration in multiple laboratories

Note 1 to entry: Throughout this document, when used in mathematical formulae, $P_{\alpha\lambda}$ refers to the estimator for the *probability of detection (POD)* (3.8) parameter across laboratories.

Note 2 to entry: The symbol for LPOD is the symbol for POD with the lowercase Greek letter ƛ (lambda) to indicate laboratory-wide.

3.6
method
procedure that includes sample processing, assay and data interpretation

3.7
naturally incurred sample
sample that contains the measurand by virtue of its inherent characteristics rather than the measurand being intentionally added

3.8
probability of detection
POD
P_{α}
probability of a positive analytical outcome of a *qualitative method* (3.9) for a given matrix at a given concentration in a single laboratory

Note 1 to entry: Throughout this document, when used in mathematical formulae, P_{α} refers to the estimator for the probability of detection parameter.

Note 2 to entry: The symbol for POD is drawn from the term P for probability and the first letter of the Greek term for detection, ανίχνευση.

3.9
qualitative method
method (3.6) of analysis with two possible outcomes

Note 1 to entry: Qualitative method is an alternative terminology to binary method.

3.10
replicate test sample
sample taken from a bulk sample such that the replicate test samples are as close to identical as achievable, in order to constitute *identical test items* (3.3)

3.11
validation experiment
determination of *method* (3.6) performance parameters from a series of test results reported by one or more usually a number of participating laboratories

3.12
upper confidence limit
UCL
$\hat{\mu}_{\mathrm{U}}$
upper value of a range containing the true value of the measurand with a specified probability

Note 1 to entry: The symbol for UCL is taken from Reference [5].

4 Characterization of a qualitative method via a validation experiment

4.1 Criteria for a standard measurement method

The following criteria should be taken into consideration when validating a qualitative method of analysis:

— applicability;

— robustness;

— selectivity;

— POD related to the measurand concentration.

All measurements shall be carried out according to a standard method based on a written document that describes in full detail how the measurement shall be carried out, including the applicability and selectivity of the method. It shall incorporate information based on the robustness testing of the method established at the single laboratory level when developing the method. The standard method may be modified by the result of experiments to determine the intermediate precision and/or the results of collaborative multi-laboratory trial(s).

4.2 Performance of a validation experiment

The estimates of performance parameters derived from a validation experiment are valid only for tests carried out according to the standard measurement method. A validation experiment can be considered to be a practical test of the adequacy of the standard measurement method. One of the main purposes of standardization is to standardize how methods are characterized, and eliminate differences between users (laboratories) as far as possible. The data provided by a validation experiment will reveal how effectively this purpose has been achieved. Pronounced differences between the laboratories often indicate that the measurement method can be improved.

From a practical point of view, it is important and desirable to carry out a number of steps before proceeding with the validation experiment. This includes: a) measurement of several replicates by one operator to establish suitable test materials that will cover the desired POD levels, followed by: b) a

mini validation experiment to establish that the instructions for the experiment are clear and sufficient and that the test materials are suitable for the full validation experiment.

4.3 Nature of test materials

Validation of qualitative methods requires the use of known positive (low and high POD) and negative (effectively as close as possible to zero POD) materials. Special challenges arise when a biological material is being tested, and pure reference material (CRM traceable back to SI units) may not be readily available. For some biomolecular methods, naturally incurred samples may be the only source of materials for validation. The preparation and source of each material shall be documented. Wherever possible, a quantitative method can be used to confirm the concentration of the measurand.

4.4 Requirements for replicate test samples

In a validation experiment, a number of replicate test samples of a specific material or specimens of a specific product are typically sent from a central point to a number of laboratories. The definition of repeatability conditions states that the measurements in these laboratories shall be performed on identical test items and refers to the moment when these measurements are actually carried out.

The test materials will ideally be evaluated for homogeneity before preparing the replicate laboratory samples to be sent to the laboratories, or by testing a number of the replicate test samples if a suitable method is available. Furthermore, the replicate test samples shall be identical test items (under the definition of ISO 5725-1) when dispatched to the laboratories and the replicate samples shall be stable and remain identical during transport and during the different time intervals that can elapse before the measurements are actually performed.

NOTE 1 The terms “identical” and “identical test items” are not the same as “identical test portions” (see ISO 5725-2:1994, Clause 5). There will always be some level of variation between replicate test samples (i.e. the actual materials sent), and this is an integral part of testing method repeatability. Test portion variability is dependent on concentration, test portion size and matrix homogeneity. When preparing the replicate test samples for a collaborative study, the concept of identical test items is to be interpreted as each test sample having an equal probability of producing a positive test result. This means that all laboratories receive essentially the same test items. The test portions will always have some level of variation, which is an inherent part of the measurement variation.

NOTE 2 The number of replicate samples required to get a good estimation (at 95 % confidence) of the LPOD for a two-sided coverage is 12 per level for the range 25 % to 75 % LPOD for the case where 8 laboratories are included (see Table E.2). If more participants are available, the number of replicate samples can be lowered in consultation with a statistician. However, the larger numbers needed to get ideal estimates of the LPOD at high and low measurand concentrations may not be practicable to achieve in a multi-laboratory trial.

Conditions should be representative of the use of the method in the laboratory. It shall be clearly stated when reporting the results if an intermediate material, such as a ground sample or an extract, is distributed for this purpose. Moreover, it shall be shown that the intermediate materials are stable under shipping conditions.

NOTE 3 While the replicate test samples supplied at each concentration would preferably consist of unprocessed material (such as whole grain or seeds) in order to test the whole method from sample to result, this is, in most cases, impractical. Therefore, it is most practical to grind the material and distribute a typical powder that would be obtained under typical conditions.

Test materials are prepared and divided into test samples before these replicate test samples are shipped to the participating laboratories. The replicate test samples may be reduced to test portions in the laboratory or analysed directly. The relationship is given in Figure 1.

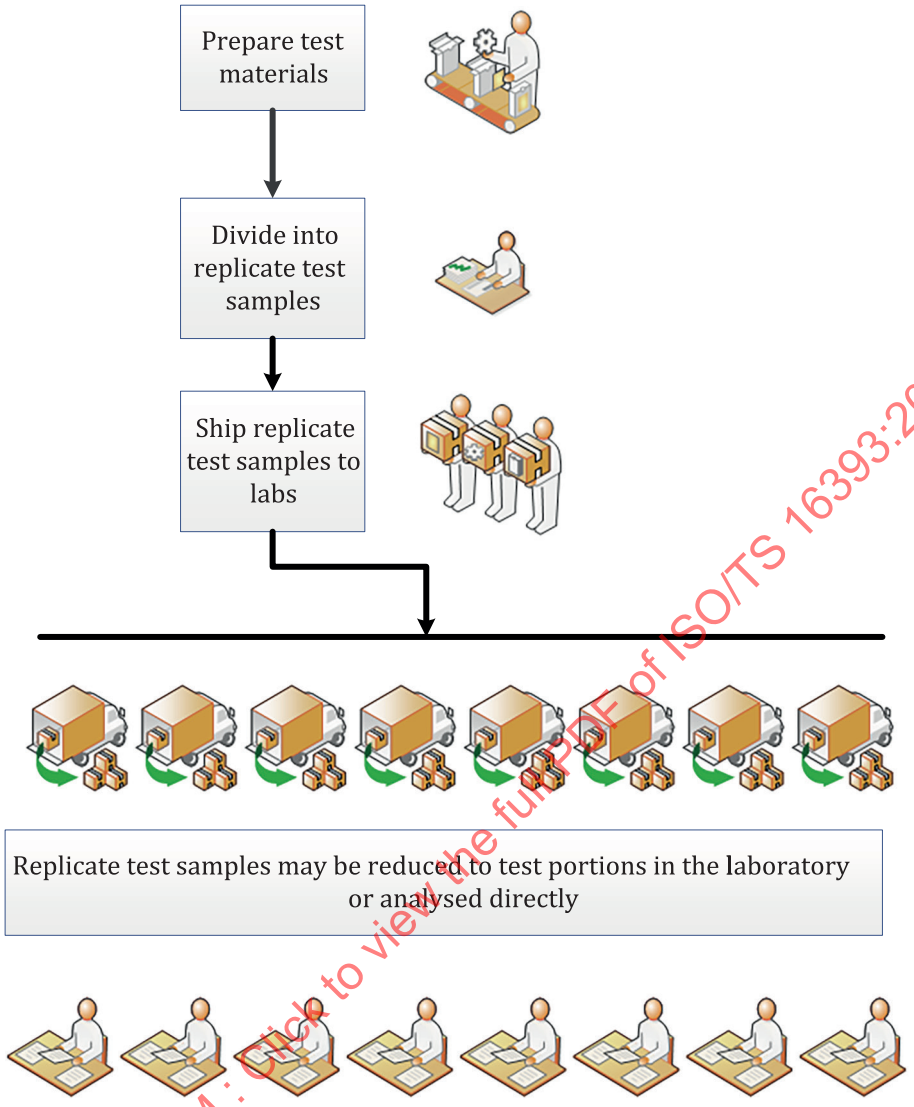


Figure 1 — Relationship between the test materials, replicate test samples and test portions

4.5 Robustness (ruggedness)

The method developer is expected to evaluate the robustness of the method against small changes in analytical conditions and external influences, and identify variables which could have a significant effect on method performance. Critical variables should be included in the standard measurement method (e.g. by including an acceptable temperature range).

4.6 Applicability

The user should be able to determine whether the method will be appropriate for the desired application (fit for purpose) and if there will be limitations to its use. Applicability is the analytes, matrices and concentrations for which a method of analysis may be used satisfactorily. An applicability statement shall therefore be provided by the method developer. It should include a list of the known analyte(s) or measurand(s) that can be determined by the method, and the form in which analyte(s) can be determined, e.g. speciation, total/available, the sample matrix(es) within which those analyte(s) can be determined. In addition to a statement of the range of capability of satisfactory performance for each factor, the statement of applicability may also include warnings as to known interference by other analytes, or inapplicability to certain matrices and situations. For example, concentrations that may lead to reduced POD at concentrations higher than those normally expected should also be specified, as

certain methods (such as those depending on antibodies) have the possibility of giving a negative result at very high concentrations of the measurand (the hook effect).

NOTE Applicability outside of the food sector can be referred to as “scope”.

4.7 Selectivity

Determination of selectivity is a single laboratory study designed to demonstrate that a method does not detect non-target measurands expected to erroneously give a positive result due to chemical or structural similarities.

The method should be shown to give a positive result for claimed measurands. Each measurand from the selectivity test panel should be tested at the appropriate target concentration for each measurand.

4.8 Experimental design for a multi-laboratory study

4.8.1 Participating laboratories

Ideally, the chosen laboratories should be a random sampling of all potential method users. Laboratories participating in any validation study for qualitative methods should have experience and training in performing the type of method being tested. However, the participating laboratories should not consist exclusively of those that have gained special experience during the process of standardizing the method. Neither should they consist (exclusively) of specialist reference laboratories, in order to demonstrate the accuracy to which the method can perform in expert hands.

Estimating the POD at applicable measurand concentrations can be carried out provided that an adequate number of replicate test samples are analysed across a suitable number of concentrations and a sufficient number of laboratories. The number of replicates per laboratory and the number of laboratories should be chosen with consideration of the effect of the size of the validation experiment on the size of the confidence intervals that will be obtained.

4.8.2 Number of laboratories

The purpose of involving a large number of laboratories in the study is to get a wider subset of potential method users to contribute data to the study. Using a large number of laboratories will reduce the subsampling error and will mean that the estimates that are obtained in the study will be less biased. In addition, with more laboratories, it will be easier to detect a laboratory effect in the data, if it is significant. The absolute minimum number of laboratories reporting and included in the final statistical analysis of the study is eight.

4.8.3 Number of levels

The minimum number of concentration levels to study should be five.

The experiment should verify that the method is sensitive to concentration, so that at low levels there is a low POD and that at a high concentration there is a POD. The experiment shall be designed to best characterize the POD curve, in as efficient a manner as possible.

One concentration level should be chosen where the expected POD is close to zero. This will demonstrate the method will not give a positive response at low, near-zero concentrations.

There should be a second concentration level where the method is expected to give > 95 % of positive responses.

There will be some concentration levels where the POD is expected to be in a marginal range (0,85 to 0,95 or 0,05 to 0,15), which is important to identify so that the response curve can be better characterized and the transition concentration from medium POD to high POD can be identified. In addition, a sample in the mid-range (35 % POD to 65 % POD) will allow the experiment to expose cases where there is a large difference in sensitivity between participating laboratories.

Alterations to the above basic scheme may be advised. Five levels would be optimal, including those in the marginal range to increase the confidence in estimation of the detection limit of the method. If the high or low POD (e.g. POD of 0,95, or 0,05) is deemed to be more important, many replicates at the high or low POD may be performed at the expense of replicates at the intermediate POD in order to focus the confidence interval of the high or low POD estimates.

4.8.4 Number of replicates per level and laboratory

In order to obtain sufficient information and maintain the required statistical confidence and accuracy of the confidence intervals, there should be at least 12 replications per level at each laboratory (for the case where there are 8 laboratories participating).

NOTE 1 The number of replicates required when characterizing a qualitative method is higher than required for quantitative methods, due to the reduced level of information provided by a qualitative method as compared to a quantitative method.

NOTE 2 Both quantitative and qualitative methods have, in practice, commonly observed systematic dependencies between mean measurand level and variance. For the qualitative case, at concentrations where the observed POD values are close to 0 or 1, very little variation will be observed in the data sets, as observations will be either mostly positive or mostly negative. At concentrations where POD values fall in the fractional range (e.g. between 0,15 and 0,85), more variation will be observed within and between the laboratories. Thus, the number of replications required by each collaborator will depend upon the range of POD of interest. Ranges approaching 0 and 1 will require more replicates as the number of positive or negative results respectively approach zero. For example, at least 12 replicates could be required for a range of 0,25 to 0,75. For a range of 0,20 to 0,80, at least 16 replicates could be required, and for a range of 0,50 to 0,90, at least 35 replicates could be desired. Thus, depending on applicability, the collaborative study could use different numbers of replicates at different concentration levels.

For PODs of 0,95 or 0,05, it would be optimal to have at least 60 replicates per laboratory, and for POD of 0,01 or 0,99, to have 300 replicates. This is not practicable in many cases. If less than the ideal number of replicates are used, the confidence intervals around the POD will be greater than if the optimum number of replicates were used. It is recognized that the recommendations stated above are sometimes not achievable in a multi-laboratory validation experiment.

The purpose of this repetition is to estimate repeatability, so these replicates should be analysed under repeatability conditions

4.9 Validation experiment under intermediate conditions

A single laboratory validation can be used when it is not practicable to carry out a full multi-laboratory collaborative study. This study should consider intermediate precision conditions, in which observations are carried out in the same laboratory, but one or more of the factors of time, operator or equipment is allowed to vary. In establishing the precision of a measurement method, it is very important to define the appropriate observation conditions, i.e. whether the above three factors should be constant or not. The statistical treatment of intermediate precision data is the same as data derived from a multi-laboratory experiment, treating multiple days, trials and/or operators as the laboratory equivalent.

4.10 Expressing the results of a validation experiment

4.10.1 General

Data from a validation experiment can be expressed in tabular form or as a graph or figure. Tables 1 and 2 show some examples for illustrative purposes.

The (tabular) data should include POD estimate and confidence interval for every level.

Table 1 — Examples of tabular results summaries for validation experiments

Conc	N	x	POD	95 % LCL	95 % UCL
0	32	1	0,031 3	0,000 0	0,157 4
0,1	320	30	0,093 8	0,066 5	0,130 7
5	320	239	0,746 9	0,696 5	0,791 4
10	320	293	0,915 6	0,880 0	0,941 4
20	320	307	0,959 4	0,931 7	0,976 1
100	32	32	1,000 0	0,892 8	1,000 0

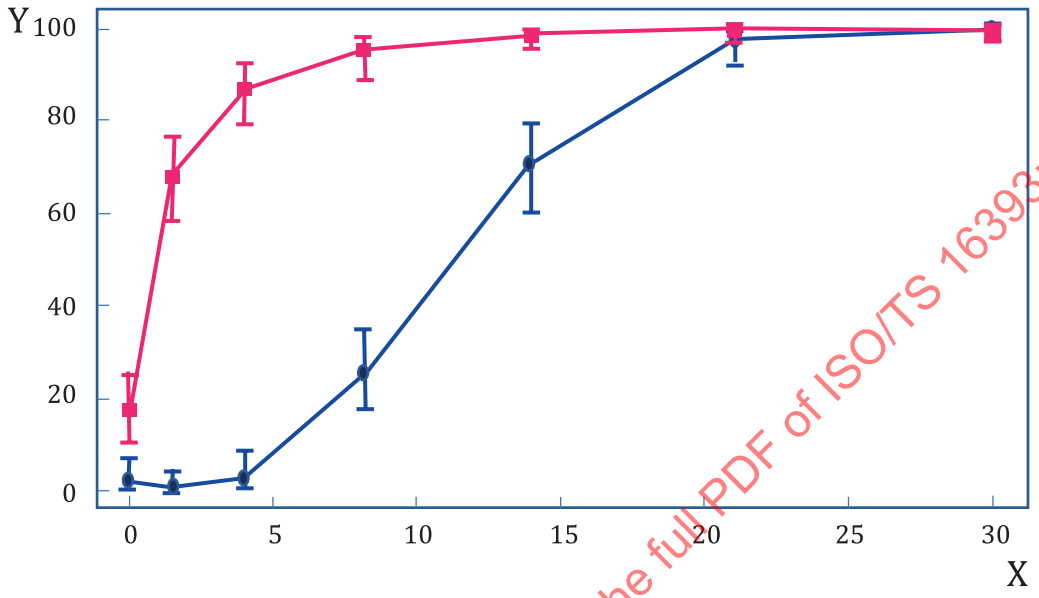
Table 2 — Example of POD values for peanut allergen test kits

	KIT A					KIT B					dPOD (A-B)		
Conc (parts per million)	**N**	**x**	**POD(A)**	**95 % LCL**	**95 % UCL**	**N**	**x**	**POD(B)**	**95 % LCL**	**95 % UCL**	**dPOD**	**95 % LCL**	**95 % UCL**
0	630	2	0,003 175	0,000 871	0,011 5	630	15	0,023 81	0,014 481	0,038 91	−0,020 63	−0,035 91	−0,008 13
1,5	630	541	0,858 73	0,829 353	0,883 759	630	601	0,953 968	0,934 672	0,967 761	−0,095 24	−0,127 69	−0,063 64
4	630	543	0,861 905	0,832 763	0,886 659	630	618	0,980 952	0,967 004	0,989 071	−0,119 05	−0,149 3	−0,090 63
8,2	630	563	0,893 651	0,867 146	0,915 384	630	626	0,993 651	0,983 789	0,997 528	−0,1	−0,126 79	−0,076 13
14	630	604	0,958 73	0,940 217	0,971 683	630	629	0,998 413	0,991 064	1	−0,039 68	−0,058 26	−0,024 79
21	630	628	0,996 83	0,988 499	0,999 129	630	630	1	0,993 939	1	−0,003 17	−0,011 5	0,003 309
30	630	630	1	0,993 939	1	630	629	0,998 413	0,991 064	1	0,001 587	−0,004 68	0,008 936

NOTE Data from Reference [18]. It shows the differences between the POD values and confidence intervals thereof.

4.10.2 Graphical representation of the data

Results of a validation experiment can be graphically presented as a plot of POD as a function of concentration (see Figures 2 and 3), with 95 % confidence intervals or via a dot plot (see Figure 4). This can be conveniently done in the R software[13], although other approaches may be possible. Details of the approaches for analysis of results of the validation experiment are described in the annex(es).



Key

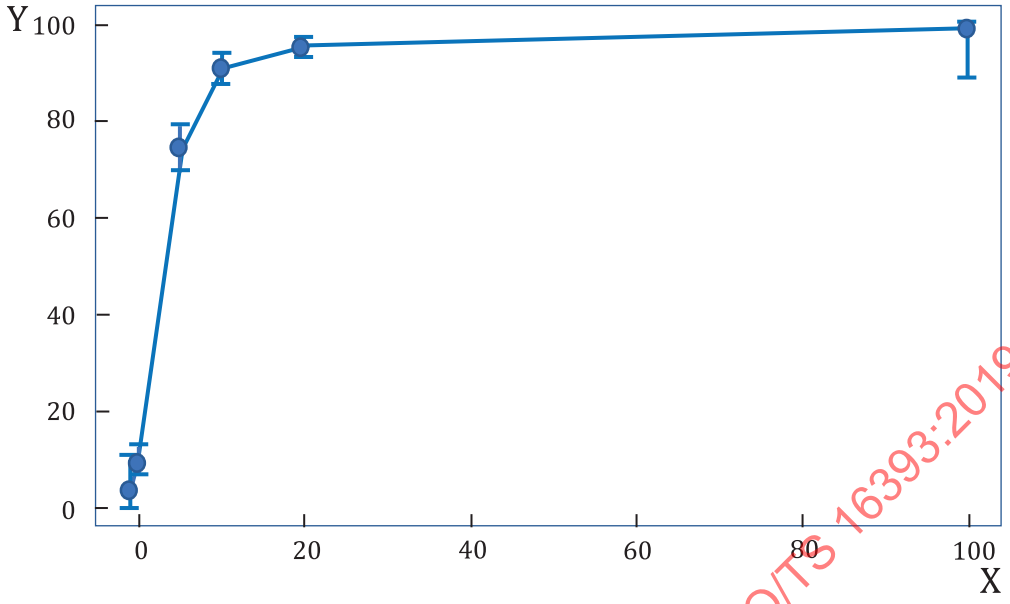
kit A	X concentration (mg/kg)
kit B	Y LPOD (%)

NOTE 1 LPOD is plotted against concentration in parts per million of peanut flour.

NOTE 2 Kit A (blue circles) has lower LPOD values than Kit B (red squares).

NOTE 3 Data from Reference [18].

Figure 2 — Examples of POD curves for two peanut allergen test kits showing 95 % confidence limits



Key

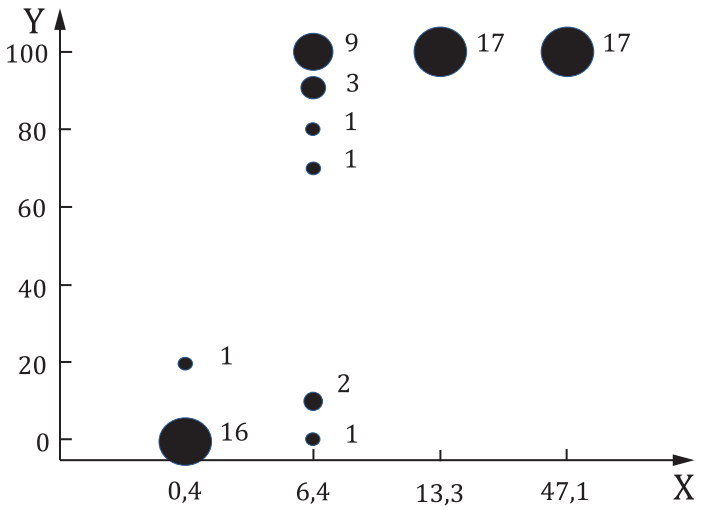
X DNA copies/aliquot

y POD (%)

NOTE 1 POD is plotted against the number of DNA copies per aliquot.

NOTE 2 Unpublished data from Reference [6].

Figure 3 — Examples of POD curves for a PCR reaction, showing 95 % confidence limits calculated according to Annex B



Key

X gluten concentration (mg/kg)

y POD (%)

NOTE 1 Number stated at each circle represents the number of laboratories with the same POD. The areas of circles are proportional to number of this number. Ten replicates were analysed per laboratory.

NOTE 2 Data from Reference [14].

Figure 4 — Percent POD observed by each of 17 participating laboratories for replicate samples with a measurand concentration of between 0,4 mg/kg and 47,1 mg/kg gluten

4.11 Calculation of the confidence interval for the general mean, confidence interval and prediction interval

Confidence limits for the general mean allow calculating limits for the POD. Confidence limits for the general mean can also be used to compare different qualitative test methods.

The raw data from all laboratories shall be analysed by a statistical model for estimating the confidence limits of estimates of the LPOD. Models that can be considered for the characterization of methods based on multi-laboratory validation based on publications and submissions by the WG experts:

— modified Wilson interval model;

  — degrees of freedom, $\nu = \Lambda - 1$ (Λ = number of laboratories) – hybrid model (see Annex B);

— maximum likelihood approaches;

  — probit model as latent variable using profile likelihood (see Annex C);

  — beta binomial as latent function using approximate likelihood (see Annex D).

These models described in Annexes B to D are applicable to cases where there is no clear scalar or concentration relationship between the concentrations of measurand in the samples provided at each POD. In cases where there is information about the scalar relationship of the samples at the PODs tested, it may be possible to use this additional information to reduce the size of the estimates of the confidence intervals, especially at the high and low PODs. Additional information will be available where the methods used are quantitative methods used to make a binary decision.

The results of using a beta binomial approach or probit approach to determine the confidence intervals of the POD may be significant in some levels of POD and not significant at other levels. The modelling experiment described in Annex E was designed to test for this situation.

NOTE At some levels, the differences between the results using a beta binomial approach and probit approach can be small enough that there is not a significant practical difference in using the two models to describe the confidence intervals.

4.12 Calculation of prediction interval for PODs in each laboratory

In addition to a confidence interval for the mean probability of a positive response across laboratories, the interval which is expected to contain the PODs of each of the laboratories that use the method is estimated. If this entire interval is inside an acceptable range for the value of the POD (e.g. if the bottom end of the interval at a particular measurand concentration is greater than the required POD at that concentration) then fit for purpose performance is expected in a sufficiently high proportion of all of the laboratories that use the method. Technically, this kind of interval is called a tolerance interval. It is analogous to intervals that describe expected measurement uncertainty for quantitative analytical methods.

This approach can also be applied to examine the effect of intermediate precision on POD. Intervals can be derived for PODs across days, operators, equipment, etc.

5 Statistical model for test result

5.1 General

The statistical model used to characterize qualitative (binary) methods is based on the POD.

5.2 Basic model

For a particular material/level combination, it is useful to assume that every test result y is the sum of three components, as shown by Formula (1):

$$y = m + B + e \tag{1}$$

where

y is the test result (limited to values 0 or 1);

m is the overall mean expected response;

B is the laboratory component of bias under repeatability conditions;

e is the random error occurring in every measurement under repeatability conditions.

5.3 Constraints in the model

In the qualitative model, there is a special case constraint for y in the binary case, as shown by Formula (2):

$$y \in \{0,1\} \tag{2}$$

In this case, with the constraint placed on y, the practical implication is that m, B and e will also be constrained for an individual replicate.

$0 \le m \le 1$

$-1 \le B \le 1$

$-1 \le e \le 1$

5.4 General mean, *m*

For quantitative methods, if *m* is in units of concentration, it is generally expected that $m = c$, where *c* is the concentration. If *m* is not a concentration (or amount) of measurand, *m* and *c* can be related by a calibration function.

For qualitative methods, this calibration cannot be easily achieved without replication, so the mean, *m*, has a special connotation in the binary model. With the coding convention (i.e. 0 = "negative" and 1 = "positive"), the mean is the mean probability of a positive response at the concentration tested. This probability is the probability of a positive response at a given concentration or POD (P_α). See Formula (3):

$$m = P_\alpha = P(+|c) \tag{3}$$

5.5 Variance parameters

Formulae (4) to (8) for the variance parameters still apply as given in the general model of ISO 5725-1:

$$\sigma_\lambda^2 = var(B) \text{ (between-laboratory variance)} \tag{4}$$

$$\sigma_W^2 = var(e) \tag{5}$$

$$\sigma_r^2 = \overline{var(e)} \text{ (repeatability variance)} \tag{6}$$

$$\sigma_r = \sqrt{\overline{var(e)}} \tag{7}$$

$$\sigma_R = \sqrt{\sigma_\lambda^2 + \sigma_r^2} \text{ (reproducibility variance)} \tag{8}$$

NOTE It can be shown that $\sigma_R^2 = m(1-m)$ [17], i.e. the reproducibility variance is directly related to the POD.

5.6 Relationship of qualitative model to the quantitative model

The qualitative model is not necessarily a separate model distinct from the quantitative model, but could be considered a special case or subset of the basic quantitative model (see ISO 5725-1).

Results of a validation experiment can be graphically presented as a plot of POD as a function of concentration, with 95 % confidence intervals. Some examples of POD curves are given in Figure 2. Variance component estimation via ANOVA with an additive model is not strictly correct for random laboratory variation adding to binary within-laboratory variation for a single replicate. However, with replication of $n \ge 12$ replicates/laboratory, the linear additive model approximates very well under the assumption of normality, e.g. see References [9] and [19], and provides for adequate parameter estimates in practice. The repeatability SD, σ_r, is estimated under the binomial model, so it is exact in the distribution sense. In the limits where POD approaches 0 or 1, observed variation is significantly

reduced so a solution needs to be developed about how underlying variance is shared between components.

5.7 Derivation of a limit of detection

The LOD is conventionally defined to be the concentration at which the method has a POD equal to 0,95 (POD = 0,95). The 5 % 0,95 point may be used for most applications.

NOTE It is generally agreed that any value of chi-square as large as 3,841 or larger is significant. Beyond this point lie only 5 % of sample values. There is only 1 chance in 20 that a random sample will have a chi-square greater than 3,841. This is called the 5 % point[15].

However, because the LOD for a method may not be well known or may require more or less certainty in specific applications (e.g. the 1 % point), POD curves can be used to estimate appropriate LODs for methods and measurands. PODs and the LODs derived from them are estimates based on a limited number of measurements. Confidence intervals, such as the 5 % point and the 1 % point, are useful to provide measurement uncertainty associated with the method, measurands and the concentrations at which they are measured. Concentrations given without significance levels should be avoided.

Annex A
(informative)

Estimation of the mean and variance

A.1 Single laboratory experiment

For N observations, code each observation y_i as shown by Formula (A.1):

$$y_i = \begin{cases} 1 \text{ if positive} \\ 0 \text{ if negative} \end{cases} \quad \text{(A.1)}$$

Let x be the total number of positive observations, as shown by Formula (A.2):

$$x = \sum y_i \quad \text{(A.2)}$$

Calculate the observed mean POD at each concentration as the ratio of the number of positives, x, to the total number tested, N, as shown by Formula (A.3):

$$P_\alpha = \hat{m} = \bar{y} = \frac{x}{N} \quad \text{(A.3)}$$

NOTE The confidence interval of the mean estimate: A method for calculating the characteristics of a method based on the data derived from a single laboratory experiment is provided. This yields estimates of the intermediate precision of the method. The number of replicates needed to provide a meaningful estimate of the confidence intervals of POD values near 0 and 1,0 will be larger than 12, in the range of 60 to 300 replicates.

A.2 Multi-laboratory experiment

For comparison of the results of laboratories, it has been proposed that the following approach be followed.

Calculation of the mean: For a multi-laboratory validation experiment, where p laboratories have submitted results, each laboratory submits n_l replicates.

The number of positive responses is shown by Formula (A.4):

$$X = \sum_{l} \sum_{i} y_{li} \quad \text{(A.4)}$$

$l = 1,2,3,\ldots p$

$i = 1,2,3,\ldots n_l$

For each level, the overall number of responses is shown by Formula (A.5):

$$N = \sum n_l \quad \text{(A.5)}$$

For each level, calculate the general mean as Formula (A.6):

$$\hat{m} = \bar{\bar{y}} = \frac{X}{N} \tag{A.6}$$

A.3 Calculation of variances of a multi-laboratory experiment

Three variances are calculated for each level. They are the repeatability variance, the between-laboratory variance and the reproducibility variance.

The repeatability variance is shown by Formula (A.7):

$$s_r^2 = \frac{\sum_{l=1}^{p}(n_l - 1)s_l^2}{\sum_{l=1}^{p}(n_l - 1)} \tag{A.7}$$

where

s_l^2 is the within-laboratory variance for the lth laboratory and is the mean squared deviation;

p is the number of laboratories;

n_l is the number of observations for the lth laboratory.

The between-laboratory variance is shown by Formula (A.8):

$$s_\lambda^2 = \frac{s_d^2 - s_r^2}{\bar{\bar{n}}} \tag{A.8}$$

where

$$s_d^2 = \frac{1}{p-1}\sum_{l=1}^{p} n_i\left(\bar{y}_i - \bar{\bar{y}}\right)^2 = \frac{1}{p-1}\left[\sum_{l=1}^{p} n_l\left(\bar{y}_l\right)^2 - \left(\bar{\bar{y}}\right)^2 \sum_{l=1}^{p} n_l\right]$$

and

$$\bar{\bar{n}} = \frac{1}{p-1}\left[\sum_{l=1}^{p} n_l - \frac{\sum_{l=1}^{p} n_l^2}{\sum_{l=1}^{p} n_l}\right]$$

Where, owing to random effects, a negative value for s_λ^2 is obtained from these calculations, the value should be assumed to be zero. The reproducibility variance is shown by Formula (A.9):

$$s_R^2 = s_r^2 + s_\lambda^2 \tag{A.9}$$

Annex B
(informative)

Hybrid modified Wilson interval model

B.1 General

The approach given in this annex (see References [3] and [19]), using transition points of LPOD = 0,15 and 0,85, and LPOD = 0,05 and 0,95, has been modelled in an improved method and can be performed using publicly available online tools.

For a multi-lab trial where Λ = number of laboratories, R = replicates per lab, $N = \Lambda R$ = total replicates, the LPOD, $P_{\alpha\lambda}$, estimate is shown by Formula (B.1):

$$P_{\alpha\lambda} = \frac{x}{N} \tag{B.1}$$

where x is the number of positive results.

B.2 Method for estimating LPOD, $P_{\alpha\lambda}$ 95 % confidence intervals

B.2.1 Step 1: Enter data into the AOAC spreadsheet with 1 for positive response and 0 for negative response. Record the mean LPOD, $s(R)$ and $s(r)$. This spreadsheet can be found at: http://www.aoac.org/imis15_prod/AOAC_Docs/NEWS/09trad04_AOAC_binary-v2-3.xls.

Alternatively, code results as 1 for positive (detected) response and 0 for negative (not detected) response. Calculate statistics as described in the ANOVA-based analysis of ISO 5725-2:1994, 7.4. Record the general mean as LPOD, $s(R)$ and $s(r)$. Do not remove any results as outliers without root cause identification.

B.2.2 Step 2: Calculate $s(\Lambda)$, standard deviation due to lab effect as shown by Formula (B.2):

$$s(\Lambda) = \sqrt{s(R)^2 - s(r)^2} \tag{B.2}$$

B.2.3 Step 3: Calculate $s(P_\alpha)$ as the standard deviation of the individual laboratory POD estimates, as shown by Formula (B.3):

$$s(P_\alpha) = \sqrt{\frac{\Sigma(P_{\alpha i} - P_{\alpha\lambda})^2}{\Lambda - 1}} \tag{B.3}$$

B.2.4 Step 4: Calculate degrees of freedom, ν, for $s(P_\alpha)$ as shown by Formula (B.4):

$$\nu = \Lambda - 1 \tag{B.4}$$

B.2.5 Step 5: Calculate 95 % confidence limits LPOD, $P_{\alpha\lambda}$, as shown by Formulae (B.5) to (B.10):

If $0{,}15 \leq P_{\alpha\lambda} \leq 0{,}85$:

$$\hat{\mu}_L = \max\left\{0, P_{\alpha\lambda} - \frac{t_{0,975,\nu}\, s\left(P_\alpha\right)}{\sqrt{\Lambda}}\right\} \tag{B.5}$$

$$\hat{\mu}_U = \min\left\{1, P_{\alpha\lambda} + \frac{t_{0,975,\nu}\, s\left(P_\alpha\right)}{\sqrt{\Lambda}}\right\} \tag{B.6}$$

If $P_{\alpha\lambda} < 0{,}15$ or $P_{\alpha\lambda} > 0{,}85$:

$$\hat{\mu}_L = \frac{x + 1{,}920\ 7 - 1{,}960\ 0\sqrt{x - \frac{x^2}{N} + 0{,}960\ 4}}{N + 3{,}841\ 5} \tag{B.7}$$

$$\hat{\mu}_U = \frac{x + 1{,}920\ 7 + 1{,}960\ 0\sqrt{x - \frac{x^2}{N} + 0{,}960\ 4}}{N + 3{,}841\ 5} \tag{B.8}$$

where

x is the number of observed positive outcomes;

N is the total number of trials.

If $P_{\alpha\lambda} = 0$:

$\hat{\mu}_L = 0$

$$\hat{\mu}_U = \frac{3{,}841\ 5}{N + 3{,}841\ 5} \tag{B.9}$$

If $P_{\alpha\lambda} = 1$:

$$\hat{\mu}_L = \frac{N}{N + 3{,}841\ 5} \tag{B.10}$$

$\hat{\mu}_U = 1$

B.3 Transition point on *x* (number positive)

The transition point can be specified by the number of positive observations, *x*. The best coverages were found when 0, 1, 2 or 3 positives (and symmetrically 93, 94, 95 and 96 positives) were used for the Wilson interval, and the Student interval was used for all other outcomes in between. Simulated coverages based on hybrid modified Wilson with Student t-interval with $\Lambda - 1$ degrees of freedom, transition point at $X = 3$ and $x = 93$ are tabulated in Annex E.

Annex C
(informative)

Maximum profile likelihood based on the probit model

Even though the observed variable *y* of the basic model (see 5.2) is binary, encoding success of detection as 1 and failure of detection as 0, the measurand itself has a (unknown) concentration. The latent variable model explicitly recognizes this fact[17]. If the concentration could be measured, the model used in ISO 5725-2 could describe the measured concentration as Formula (C.1):

$$X = m + B + e \tag{C.1}$$

where

X is the measured concentration;

m is the true concentration;

B is the laboratory bias;

e is a random measurement error.

Both B and e are assumed to be random and to follow a normal distribution with mean zero and variances σ_R^2 and σ_r^2, respectively. A positive detection with $y = 1$ in the basic model is then given if and only if $X > 0$ in the latent variable model.

The POD, P_α, is then modelled by a probit-link, given by the LPOD, $P_{\alpha\lambda}$, as shown by Formula (C.2):

$$P_{\alpha\lambda} = P(X > 0) = P(B + e > -\mu) = \Phi(\mu + B) \tag{C.2}$$

where Φ is the cumulative distribution function of the standard Normal distribution.

Thus, the probability of detection in the basic model is translated into the probability that the latent variable describes a positive concentration of the measurand. The link between the concentration and the detection probability is given by the probit, which assumes that the random number of positive detections of a laboratory follows a binomial distribution with laboratory-specific POD equal to $\Phi(\mu + b)$; here b is again the laboratory bias.

The parameters μ and $\sigma = \sigma_£$ for the mean and the laboratory standard deviation are calculated using a profile-likelihood approach maximizing the log-likelihood, as shown by Formula (C.3):

$$\ln £(\mu,\sigma) = \sum_{i=1}^{k} \ln P(Y_i = n_i \mid \mu,\sigma) \tag{C.3}$$

with

$$P(Y_i = n_i \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} g_n\left(n_i, \Phi(\mu + b)\right) \exp\left(-\frac{b^2}{2\sigma^2}\right) db$$

where g_n (x,p) denotes the probability density function of the binomial distribution with parameters n and p.

No closed-form solution to this maximization problem exists, but numerical solution of this problem is standard and a web service is available from Reference [12].

Using the maximum-likelihood estimates, £, for the mean and variance, the LPOD is then estimated as Formula (C.4):

$$P_{\alpha\lambda} = \Phi\left(\frac{\mu_{£}}{\sqrt{\sigma_{£}^2 + 1}}\right) \tag{C.4}$$

Lower and upper confidence limits of the 95 % confidence interval are calculated by Formulae (C.5) and (C.6):

$$\hat{\mu}_{L} = \min\left\{\Phi\left(\frac{\mu}{\sqrt{\sigma^2 + 1}}\right); \ln £(\mu_{£}, \sigma_{£}) - \ln £\left(\mu, \sigma^2\right) \leq 0{,}5\, t^2_{0{,}975,\, k-1}\right\} \tag{C.5}$$

and

$$\hat{\mu}_{U} = \max\left\{\Phi\left(\frac{\mu}{\sqrt{\sigma^2 + 1}}\right); \ln £(\mu_{£}, \sigma_{£}) - \ln £\left(\mu, \sigma^2\right) \leq 0{,}5\, t^2_{0{,}975,\, k-1}\right\} \tag{C.6}$$

Annex D
(informative)

Maximum likelihood estimate based on beta binomial distribution

A normal distribution has a definition range of (-∞, ∞) but the POD is bounded within [0, 1]. In estimating the fluctuation of such variable, a beta-distribution is frequently used instead of a normal distribution, e.g. References [10], [20] and [21]. A beta-distribution has a definition range [0, 1] and it is usually sufficiently flexible, although it has only two parameters as well as a normal distribution. Let $P_{\alpha i}$ be the POD in the ith laboratory (i = 1,2,3,...,m) . Then, a beta-distribution can be used to describe the variability of $P_{\alpha i}$ among laboratories. See Formula (D.1):

$$f(p)=\frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1} \tag{D.1}$$

where

a and b are positive constants;

$B(.,.)$ is a beta function.

The mean, $P_{\alpha 0}$, and variance, $V(P_{\alpha i})$, are given by $P_{\alpha 0}=a/(a+b)$ and $V(P_{\alpha i})=ab/\left[(a+b)^2(a+b+1)\right]$, respectively. Let X_i and N_i be the number of detected experiments and the number of total experiments in the ith laboratory, respectively. A binomial distribution can be used for X_i. See Formula (D.2):

$$Pr(X_i=x_i|P_{\alpha i}=p,N_i=n_i)=\binom{n_i}{x_i}p^{x_i}(1-p)^{n_i-x_i} \tag{D.2}$$

Then, the probability density of X_i is given by a beta-binomial distribution. See Formula (D.3):

$$Pr(X_i=x_i|N_i=n_i)=\int_0^1 Pr(X_i=x_i|P_{\alpha i}=p,N_i=n_i)f(p)dp=\binom{n_i}{x_i}\frac{B(a+x_i,b+n_i-x_i)}{B(a,b)} \tag{D.3}$$

The mean and variance are given by $n_i a/(a+b)$ and $n_i ab(a+b+n_i)/[(a+b)^2(a+b+1)]$, respectively. The total logarithmic likelihood is given by Formula (D.4):

$$\sum_{i=1}^{m}\left(\log_e Pr(X_i=x_i|N_i=n_i)\right) \tag{D.4}$$

The parameter b in Formula (D.3) can be replaced with $a\left((1/P_\alpha)-1\right)$. Then, the maximum likelihood estimates of parameters, $P_{\alpha 0}$ and a, are obtained by maximizing the quantity of Formula (D.4). The logit scale and the logarithmic scale should be used in finding the optimal $P_{\alpha 0}$ and optimal a, respectively, to improve the property of maximum likelihood estimates. The asymptotic confidence interval is calculated from the variance-covariance matrix estimated by the inverse of Hessian matrix.

The prediction interval, instead of the tolerance interval, is calculated by the quantile of Formula (D.1) after substituting the maximum likelihood estimates. A negative-binomial distribution should be used