PRE-STUDY ANALYTICAL METHOD VALIDATION: COMPARISON OF FOUR ALTERNATIVE APPROACHES BASED ON QUALITY LEVEL ESTIMATION AND TOLERANCE INTERVALS

GOVAERTS, B., DEWE, W., MAUMY, M. and B. BOULANGER

IAP STATISTICS NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP
Pre-study analytical method validation: comparison of four alternative approaches based on quality level estimation and tolerance intervals

Bernadette Govaerts\textsuperscript{a1}, Walthère Dewé\textsuperscript{b}, Myriam Maumy\textsuperscript{c} and Bruno Boulanger\textsuperscript{d}.

\textsuperscript{a} Université Catholique de Louvain, Institut de Statistique, Belgium
\textsuperscript{b} GlaxoSmithKline Biologicals, Rixensart, Belgium
\textsuperscript{c} Université Louis Pasteur, Laboratoire de Statistique, France
\textsuperscript{d} UCB Pharma SA, Braine-L’alleud, Belgium

SUMMARY

In industry and in laboratories, it is crucial to continuously control the validity of the analytical methods used to follow the products’ quality characteristics. Validity must be assessed at two levels. The “pre-study” validation aims at demonstrating before use that the method will be able to achieve its objectives. The “in-study” validation is intended to verify, by inserting QC samples in routine runs, that the method remains valid over time. At these two levels, the analytical method will be claimed valid if it is possible to prove that a sufficient proportion of analytical results is expected to lie within given acceptance limits $[-\lambda, \lambda]$ around the nominal value.

This paper presents and compares four approaches to checking the validity of a measurement method at the pre-study level. They can be classified into two categories. In the first, a lower confidence bound for the estimated probability $\pi$ of a result lying within the acceptance limits is computed and compared to a given acceptance level. Maximum likelihood and delta methods are used to estimate the quality level $\pi$ and the corresponding estimator variance. Two approaches are then proposed to derive the confidence bound: the asymptotic maximum-likelihood approach and a method proposed by Mee \cite{1}. The second category of approaches checks whether a tolerance interval for hypothetical future measurements lies within the predefined acceptance limits $[-\lambda, \lambda]$. $\beta$-expectation and $\beta$-content tolerance intervals are investigated and compared in this context.

These four approaches are illustrated on a bioanalytical HPLC-UV analytical process and compared through simulations.

\textsuperscript{1} Correspondance to: Institut de Statistique, Université Catholique de Louvain, 20 Voie du Roman Pays, 1348 Louvain-la-Neuve, Belgique. Tél: 00-32-10474313. E-mail: bernadette.govaerts@uclouvain.be.
In industries and research laboratories, analytical methods are used to keep an eye on the material developed and produced. If the quality of an analytical method is doubtful, then the whole set of decisions based on those measures is questionable. For this reason, being able to assess the quality or “validity” of an analytical method is far more than a statistical challenge; it is a matter of ethics and good business practices.

The validity of an analytical method must be assessed at two levels. The “pre-study” validation aims at showing, by an appropriate set of experiments, that the method is able to achieve its objectives. The “in-study” validation is intended to verify, by inserting quality control (QC) samples in routine runs, that the method remains valid over time. The later will not be covered here, even if similar concepts also apply.

This paper focuses on procedures to achieve the ultimate goal of a pre-study validation, i.e. checking that the method can routinely produce a sufficient proportion of analytical results lying within acceptance limits, say $[-\lambda, \lambda]$, around the (unknown) nominal value. Indeed, several aspects of the analytical method must be studied at a pre-study level. As Hubert et al. [2] point out, a validation protocol should demonstrate the specificity, validate the response function, validate the quantification limit, validate the range, assess the linearity of results, and when all these factors have been checked, estimate the precision and trueness of the analytical method. The precision and trueness are usually studied through a validation experiment aimed at proving that, within the range of variation of the method, it is able to provide accurate results.

This paper presents and compares four approaches to assessing the validity of a measurement method at the pre-study level. They can be classified into two categories. In the first, a lower confidence bound for the probability $\pi$ of a result lying within the acceptance limits is computed and compared to a given acceptance level. Maximum likelihood and delta methods are used to estimate the quality level $\pi$ and the corresponding estimator variance. Two approaches are then proposed to derive the confidence bound: the asymptotic maximum likelihood approach and a method proposed by Mee [1]. The second category checks whether a tolerance interval for hypothetical future measurements lies within the predefined acceptance limits $[-\lambda, \lambda]$. $\beta$-expectation and $\beta$-$\gamma$ content tolerance intervals are investigated and compared in this context.

These methods are not all new; some have already been introduced, for example, in [2,3,4]. However, this paper presents those methods in a common framework, generalises former results to very general validation experimental designs, and compares them on a case study and through simulations. Their respective characteristics will be emphasised: statistical power, interpretability, ease of use, limitations, etc.

Section 2 defines the main goal of method validation and reviews the way this goal is achieved in the classical literature. Experimental designs and statistical models linked to validation experiments are also presented and a case study illustrates classical validation rules. Section 3 presents four recent approaches to method validation in a very general framework. The formulae provided can be applied to any balanced factorial validation experiment (and variance component model) and are, in particular, reformulated in the simple ANOVA I random model. The case study gives a first illustration and comparison of these four approaches. Finally, Section 4 compares the four approaches more systematically on the basis of simulations.
2. PRE-STUDY METHOD VALIDATION: PRINCIPLES AND CLASSICAL APPROACHES

2.1 Principles and vocabulary

The objective of any analytical method is to quantify accurately each of the unknown quantities that the laboratory will have to determine. In other words, the analytical method is expected to give results (Xs) for which the difference from the unknown “true” value ($\mu_T$) of the sample is sufficiently small, for example less than a predefined acceptance limit, $\lambda$, i.e.

$$-\lambda < X - \mu_T < \lambda \iff |X - \mu_T| < \lambda.$$ 

Two components may influence this difference: the bias (or trueness) of the method, and its precision. Classical method validation tools usually check the size of these two components separately but this approach has the drawbacks that, first, a very low value of one component may not compensate for a large value of the other and, second, it puts the focus on the performance of the analytical method while it should be on the results and their accuracy.

The approaches presented in this paper consider the problem globally in defining an analytical method as acceptable if it is “very likely” that the difference between any sample measurement X and its “true” value ($\mu_T$) is within the acceptance limits $[-\lambda, +\lambda]$ predefined by the analyst. The notion of “very likely” can be translated to the probabilistic equation:

$$\pi_{\text{min}} \geq P\left(\frac{|X - \mu_T|}{\sigma_X} < \lambda \right),$$

where $\pi_{\text{min}}$ is the acceptance level, and $\pi$ the quality level. The value of $\lambda$ must be chosen in agreement with the intended use of the results. The probability $\pi_{\text{min}}$ must also be fixed by the analyst, according to cost, consumer and analytical domain requirements. A key is to ensure coherence between the $\pi_{\text{min}}$ and $\lambda$ values targeted in the pre-study and in-study phases [4].

Under the assumption of normality of the measurement results $X \sim N(\mu, \sigma_X^2)$, it is easy to establish the relationship between the quality level $\pi$, the bias (or systematic error) $\delta = \mu - \mu_T$, and the precision (random error) $\sigma_X$ as

$$\pi = \int_{-\lambda}^{+\lambda} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz = P\left(\left|\frac{X - \mu_T}{\sigma_X}\right| < \lambda\right),$$

where $Z$ is a standard normal random variable. This leads to a definition of the “acceptance region”, i.e. the set of $(\delta, \sigma_X)$’s such that the quality level $\pi$ is greater than $\pi_{\text{min}}$. Figure 1 shows, below the curves, the acceptance region for various values of $\pi_{\text{min}}$ (99%, 95%, 90%, 80% and 66.7%) when the acceptance limits are fixed at $[-15\%, +15\%]$ as recommended by FDA [5] for bioanalytical methods. Note that, in this graph, $\delta$ and $\sigma_X$ must be interpreted as relative bias and relative standard deviation.
Before releasing an analytical method for routine use on unknown samples, it is quality and regulatory practice to perform an appropriate set of experiments to evaluate whether it meets the criteria described above. These experiments allow to estimate the bias $\delta$ and the precision $\sigma_X$, the unknown intrinsic performance parameters of the analytical procedure and then to evaluate whether, given these estimates obtained, the proportion $\pi$ of measures of new unknown samples that fall within the acceptance limits is greater than the predefined acceptance level $\pi_{min}$ (see Equation (1)).

### 2.2 Design of a validation study and related statistical models

The validation experiment must mimic as closely as possible what will happen in the routine use of the method in order to estimate related sources of variation. Common approaches consider two sources of variation: the pure method-repeatability variance and a between-run variance which takes into account a second source of variance (such as the recalibration of the measurement device, a change in the reagent batch, the machine or the operator). However, a validation experiment may also try to quantify several such components simultaneously. In all cases, repeated crossed or nested factorial-balanced designs are recommended and usually employed.

In order to estimate the total (intermediate or reproducibility) variance $\sigma^2_X$ of the measurement method, a variance-components model should be fitted to the experiment results. If $F_1, F_2, \ldots, F_{p-1}$ denote the $p-1$ variance-components factors and $\epsilon$ the repeatability factor, the general form of such a model can be written as given by Searle et al. [6] as

$$X_{i_1, i_2, \ldots, i_p} = \mu + F_{i_1} + F_{i_2} + \ldots + F_{i_{p-1}} + \epsilon_{i_1, i_2, \ldots, i_p} \quad \text{with} \quad F_{i,j} \sim iN(0, \sigma^2_j) \quad \text{and} \quad \epsilon \sim iN(0, \sigma^2_{\epsilon})$$

The parameters of this model can be estimated by classical analysis of variance for balanced designs, and by restricted maximum likelihood for unbalanced ones.

For balanced cases, the estimator of the method mean $\mu$ is

$$\hat{\mu} = \bar{X} = \frac{1}{N} \sum_{i_1, i_2, \ldots, i_p} X_{i_1, i_2, \ldots, i_p}$$

where $N$ is the total number of experiments. The total variance $\sigma^2_X$ is then defined as

![Figure 1: Acceptance region of an analytical method as a function of the method bias and precision when $\lambda = 15\%$.](image)
\[ \sigma_x^2 = \sum_{j=1}^{p-1} \sigma_j^2 + \sigma_\epsilon^2 \]

and, for balanced designs, estimated as a weighed sum of factors’ mean squares,

\[ \hat{\sigma}_x^2 = \sum_{j=1}^{p-1} \hat{\sigma}_j^2 + \hat{\sigma}_\epsilon^2 = \sum_{j=1}^{p} c_j MS_j \]

where the \( c_j \)'s depend on the model and design structures [6].

In the particular case of a one-way random model with \( n_a \) runs and \( n_w \) replications within each run, the model becomes

\[ X_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{with} \quad \alpha_i \sim iN(0, \sigma_\alpha^2) \quad \text{and} \quad \epsilon_{ij} \sim iN(0, \sigma_\epsilon^2). \]

The parameters are estimated as

\[ \bar{X} = \frac{1}{N} \sum_{i=1}^{n_a} \sum_{j=1}^{n_w} X_{ij} \quad \hat{\sigma}_x^2 = \hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2 = \frac{1}{n_w} MSA + \left(1 - \frac{1}{n_w}\right) MSE \]

where

\[ MSA = \frac{n_w}{n_a - 1} \sum_{i=1}^{n_a} (\bar{X}_i - \bar{X})^2 \quad MSE = \frac{1}{n_a(n_w - 1)} \sum_{i=1}^{n_a} \sum_{j=1}^{n_w} (X_{ij} - \bar{X}_i)^2 \]

### 2.3 Classical validation rules

Conventional literature on method validation approaches the question through the separate evaluation of several performance criteria among which trueness (bias) and precision to finally decide whether the method will be acceptable for routine use while it’s the ability of the method to provide accurate results that should be accepted. Simply, it is assumed that, if a method has acceptable performances, it will provide accurate results. This is not always correct. However, accurate results can only be obtained with appropriate methods.

For example, the FDA guidance for industry [5,7] recommends verifying, on the basis of experiments, that the bias and (intermediate or reproducibility) precision are each lower than an acceptance limit usually fixed as a percentage of the concentration, \( \lambda \) say. Such a rule may be written as:

**FDA rule:** Accept the method if: \( \hat{\delta} = |\hat{\mu} - \mu_r| < \lambda \) and \( \hat{\sigma}_x < \hat{\lambda} \).

Figure 2a shows that this acceptance rule does not respond to the requirements of Equation (1) as its acceptance region is a rectangle. Such a rule may disallow situations were the bias is low and the precision is high, which may lead to the method being rejected even when it has a reasonable quality level. On the other hand, a method with medium bias and precision will be accepted even when the quality level is low.
Figure 2: Comparison of the acceptance regions generated by classical method-validation acceptance rules ((a) FDA rule, (b) total error rule, (c) loss function rule)

Lee et al. [8] improved this approach by introducing the concept of total error. They consider a method as valid if the sum of absolute bias and precision, called the total error, is smaller than a defined acceptance level $\lambda$. Their acceptance rule can be written as:

**Total error rule:** Accept the method if $\left|\hat{\delta}\right| + \hat{\sigma}_X = \left|\hat{\mu} - \mu_r\right| + \hat{\sigma}_X < \lambda$

The shape of the corresponding acceptance region, reproduced in Figure 2b, is closer to the requirements of Equation (1). It could be improved even further by defining the total error as the square root of the sum of squares of bias and precision:

**Loss function rule:** Accept the method if $\sqrt{\hat{\delta}^2 + \hat{\sigma}_X^2} = \sqrt{(\hat{\mu} - \mu_r)^2 + \hat{\sigma}_X^2} < \lambda$

This rule is close to a loss function concept, has better statistical properties and an acceptance region very close to the ideal (Figure 2c). However it does not seem to be mentioned in the literature, although Lee et al. [8] do introduce the interesting concept of a total error profile which is close to the quality level and accuracy profiles described in Section 3 below.

All these rules have two drawbacks. First, they do not include any result-oriented direct control of the quality level $\pi_{\text{min}}$ as they only work on method performance criteria to be related one way or the other to an acceptance level $\lambda$ that is however linked to the results. Second, they do not take any account of the uncertainty in the estimators of bias and precision (which may be quite high when the number of runs of the validation experiment is small).

Indeed, the acceptance regions in grey in Figure 2 are drawn assuming that the TRUE relative bias and precision are known. However the rules are based on the observed or estimated relative bias and precision, and as a consequence the risk to accept methods truly outside those grey regions may be high. Also a method truly inside the grey region could likely be rejected while valid. In practice, an acceptance region based on estimated performance parameters must be (much) smaller (depending on risk and sample size) and totally included within the grey regions to ensure little chance to accept a method that is not valid and, as a result, protect the laboratory client.

The method-validation procedures presented in this paper, and already partially available in the literature [2, 3, 4], aim at improving these “classical” approaches by directly focusing on the validation principles described in Section 2.1 above.
2.4 Case study

Validation data from an HPLC-UV bioanalytical method [9] will be used to explore the problem described above. The design consisted of three runs with four replications per run at four concentration levels. As in routine use, the calibration model was re-estimated for each run on the basis of additional calibration data. An ANOVA I random model was fitted to the validation data for each concentration, and the corresponding bias $\hat{\delta}$ and total variance $\hat{\sigma}^2$ were estimated.

The validation data are represented graphically in terms of relative bias in Figure 3 and are all between the acceptance limits $[-\lambda, +\lambda]$ set here at +/-15% of the true value. The aim is to show that a proportion $\pi_{min}=0.8$ of future measurements can be expected to lie within these limits.

Table 1 provides, for each of the four concentration levels, the estimates of the bias and the total variance, and the results necessary to apply the classical validation rules (i.e. the (relative) bias, the precision and the total errors less than 15%). The method validity is then accepted with respect to classical validation rules.

![Figure 3: HPLC-UV validation data](image)

**Table 1:** HPLC-UV validation experiment: data necessary for the application of the classical validation rules

| Concentration | $\mu_r$ | $\bar{X}$ | $\hat{\delta}$ | $|\hat{\delta}|/\mu_r$ | $\hat{\sigma}_x$ | $\hat{\sigma}_x/\mu_r$ | $|\hat{\delta}|+\hat{\sigma}_x$ | $(|\hat{\delta}|+\hat{\sigma}_x)/\mu_r$ | $\sqrt{\hat{\delta}^2+\hat{\sigma}_x^2}/\mu_r$ |
|---------------|--------|---------|---------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 1             | 25.35  | 25.33   | -0.02         | 0.1%           | 1.61           | 6.4%           | 1.63           | 6.4%           | 6.40%          |
| 2             | 48.24  | 46.35   | -1.89         | 3.9%           | 2.39           | 5.0%           | 4.28           | 8.9%           | 6.34%          |
| 3             | 437.82 | 420.96  | -16.86        | 3.8%           | 22.11          | 5.1%           | 38.97          | 8.9%           | 6.36%          |
| 4             | 838.65 | 850.76  | 12.11         | 1.4%           | 50.80          | 6.1%           | 62.91          | 7.5%           | 6.26%          |
3. VALIDATION RULES BASED ON QUALITY-LEVEL ESTIMATION AND TOLERANCE INTERVALS

This section presents four validation rules which aim to prove the inequality

$$\pi = P(-\lambda < X - \mu_r < \lambda) \geq \pi_{\text{min}}$$

on the basis of the validation experiment. However, there is neither an exact solution nor an easy way to demonstrate its validity, even for very simple experimental designs [4]. Two classes of approximate approaches are discussed below. The first consists of calculating the lower limit of a one-tailed confidence interval for \(\pi\), and the second is based on the notion of a tolerance interval. For each class, two alternative solutions are proposed, leading to four possible ways of evaluating the validity of an analytical method.

3.1 Quality-level estimation rules

The first approach, already introduced for simple cases by Boulanger et al. [4], consists of deriving from \(\hat{\delta}\) and \(\hat{\sigma}_x\) the lower bound of a one-tailed confidence interval on the quality level \(\pi\) and checking whether or not this bound is larger than \(\pi_{\text{min}}\). However, such a probability it is not easy to estimate. No exact solution exists even with the simplest sampling scheme. Nevertheless, a variety of statistical approaches may be used to approximate a solution.

Two approaches are presented here. Both are based on the same principle: the quality level \(\pi\) is first estimated by maximum likelihood and an asymptotic estimator of the variance \(V(\hat{\pi})\) for this estimator is derived by the delta method. The maximum likelihood estimator of \(\pi\) can be derived, through the invariance principle, from the maximum likelihood estimators of the parameters of the variance components model,

$$\hat{\pi} = \phi \left( \frac{\lambda - \hat{\delta}}{\hat{\sigma}_x} \right) - \phi \left( \frac{-\lambda - \hat{\delta}}{\hat{\sigma}_x} \right) \text{ with } \hat{\delta} = \hat{\mu} - \mu_r$$

where \(\phi\) denotes the distribution of a standard normal variable. To simplify further generalisations, the RMLE unbiased estimator will be taken for \(\hat{\sigma}_x^2\) instead of the classical maximum likelihood one.

For balanced data, \(\hat{\delta}\) and \(\hat{\sigma}_x^2\) are independent and \(V(\hat{\pi})\) can be approximated asymptotically through the delta method by:

$$\text{Var}(\hat{\pi}) \cong \left( \frac{\partial \hat{\pi}}{\partial \hat{\delta}} \right)^2 V(\hat{\delta}) + \left( \frac{\partial \hat{\pi}}{\partial \hat{\sigma}_x^2} \right)^2 V(\hat{\sigma}_x^2) = \frac{1}{\sigma_x^2} \left( \phi \left( \frac{-\lambda - \hat{\delta}}{\hat{\sigma}_x} \right) - \phi \left( \frac{\lambda - \hat{\delta}}{\hat{\sigma}_x} \right) \right)^2 V(\hat{\delta})$$

$$+ \frac{1}{4\sigma_x^2} \left( -\lambda - \hat{\delta} \phi \left( \frac{-\lambda - \hat{\delta}}{\hat{\sigma}_x} \right) - (\lambda - \hat{\delta}) \phi \left( \frac{\lambda - \hat{\delta}}{\hat{\sigma}_x} \right) \right)^2 V(\hat{\sigma}_x^2)$$

where \(\phi\) denotes the density function of a standard normal variable. For balanced designs, the variances of \(\hat{\delta}\) and \(\hat{\sigma}_x\) can be estimated from the factor mean squares as
\[ V(\hat{\delta}) \doteq \sum_{j=1}^{p} h_j MS_j \text{ and } V(\sigma_X^2) \doteq \sum_{j=1}^{p} c_j^2 \frac{2MS_j^2}{r_j} \]

where the \( h_j \), \( c_j \)s and \( r_j \)s depend on the experimental design [6]. An estimator \( \hat{V}(\hat{\delta}) \) for \( Var(\hat{\delta}) \) can be derived by replacing \( V(\hat{\delta}) \) and \( V(\sigma_X^2) \) by these estimators in the above formula.

In the ANOVA I sampling scheme these variances become

\[ V(\hat{\delta}) \doteq \frac{MSA}{N} \text{ and } V(\sigma_X^2) \doteq \frac{2}{n_w^2} \left( \frac{1}{n_\alpha - 1} MSA^2 + \frac{n_w - 1}{n_\alpha} MSE^2 \right). \]

A lower bound for a confidence interval for \( \pi \) is then derived on the basis of an approximate distribution of the estimator \( \hat{\pi} \). The maximum likelihood approach supposes that \( \hat{\pi} \) is asymptotically normal and the lower bound for \( \pi \) is then defined as

\[ \pi_{\text{ML}}^{\inf} = \hat{\pi} - z_{1-\alpha} \sqrt{\hat{V}(\hat{\pi})}, \]

where \( z_{1-\alpha} \) is the \((1-\alpha)\) percentile of a N(0,1) variable.

The second approach, introduced by Mee [1], supposes that \( n' \pi \) can be approximated by a binomial distribution \( Bi(n', \pi) \). Some tedious calculations lead to the following formula for the confidence interval lower bound:

\[ \pi_{\text{Mee}}^{\inf} = \frac{n'}{n' + \frac{z_{1-\alpha}^2}{2n'}} \left( \frac{z_{1-\alpha}^2}{4n' \pi (1-\pi)} + \frac{1}{n'} (1-\pi) \hat{\pi} \right) \]

with \( n' = \frac{\hat{\pi}(1-\hat{\pi})}{\hat{V}(\hat{\pi})} \).

### 3.2 Tolerance-interval rules

A tolerance interval is defined as a “statistical interval in which, with some confidence, a specified proportion of the population will fall”. The application of this concept to analytical method validation is quite natural: if a tolerance interval of level \( \pi_{\text{min}} \) is calculated from validation data for future measurements, the interval should be included within the acceptance interval \([\mu_T - \lambda, \mu_T + \lambda]\) for a valid method. If the model parameters are known, this interval is defined as:

\[ [\mu - z_{1-\pi_{\text{min}}} \sigma_X, \mu + z_{1-\pi_{\text{min}}} \sigma_X]. \]

When the parameters are unknown and replaced by their estimators, the tolerance interval width must be adapted to take the uncertainty in both estimators into account. This idea has already been introduced in the framework of method validation [2, 3, 4]. These authors propose the \( \beta \)-expectation and the \( \beta-\gamma \) content confidence intervals as possible solutions. These are defined in general terms by Mee [10] as:

\[ E_{\delta, \sigma_X} \left\{ P_X \left( \delta - k_E \sigma_X < X - \mu_T < \delta + k_E \sigma_X \mid \hat{\delta}, \hat{\sigma}_X \right) \right\} = \beta, \]

and

\[ P_{\delta, \sigma_X} \left\{ P_X \left( \delta - k_C \sigma_X < X - \mu_T < \delta + k_C \sigma_X \mid \hat{\delta}, \hat{\sigma}_X \right) > \beta \right\} = \gamma \]

respectively, where \( \beta \) is set at the acceptance level \( \pi_{\text{min}} \) in this context and \( \gamma \) is typically set at 0.5 to give a similar result to the \( \beta \)-expectation interval. The values of \( k_E \) and \( k_C \) depend on the
experimental design used for validation and cannot be derived exactly if the design is more complicated than a simple sampling scheme.

From [3, 6, 10, 11] it can be shown that, for the general variance-component model, $k_E$ and $k_C$ can be approximated by

$$
k_E = \frac{t}{\sqrt{\frac{1}{N_E}} + \frac{1}{N_E}} \quad \text{and} \quad k_C = \frac{z_{\pi_{\min}}}{C} \frac{1}{\sqrt{\frac{1}{N_E}} + \frac{1}{N_E}},$$

where $N_E$ is the effective sample size taking into account the uncertainty of the mean $\mu$ and is estimated as

$$N_E \approx \frac{\hat{\sigma}_x^2}{\hat{\mu}}.$$

The uncertainty on $\sigma_x$ is taken into account differently in the two intervals. In the $\beta$-expectation tolerance interval, the $z$ percentile is replaced by a percentile of a $t$ distribution with $f$ degrees of freedom. For balanced data, $f$ can be estimated through the Satterthwaite formula [12] as

$$f \approx \frac{\left( \sum_j c_j^2 MS_j \right)^2}{\sum_j c_j^2 MS_j^2 r_j}.$$

In the $\beta-\gamma$ content tolerance interval, $\sigma_x$ is replaced by the upper bound of a confidence interval of level $(1-\gamma)$ on $\sigma_x$. This is obtained by multiplying $\sigma_x$ by the correction factor $C_\sigma$ defined by Graybill [11], for balanced data, as

$$C_\sigma = \frac{1}{\hat{\sigma}_x^2} \sqrt{\left( \sum_j c_j^2 H_j^2 MS_j^2 \right)}$$

with $H_j = \frac{1}{F_{\gamma, r_j, \infty}} - 1$.

For the ANOVA I model, these quantities become

$$N_E = \frac{n_a (MSA + (n_w - 1)MSE)}{MSA}, \quad f = \frac{(MSA + (n_w - 1)MSE)^2}{\frac{1}{n_a - 1} MSA^2 + \frac{n_w - 1}{n_n} MSE^2},$$

and

$$C_\sigma = \sqrt{1 + \frac{1}{\hat{\sigma}_x^2} \left( H_1^2 \left( \frac{1}{n_w} MS_A^2 + H_2^2 \left( 1 - \frac{1}{n_n} \right) MS_E^2 \right) \right)}$$

with $r_1 = n_a - 1$ and $r_2 = n_n (n_w - 1)$.

### 3.3 HPLC-UV case study

Table 2 and Figure 4 show the application of the four validation approaches to the HPLC-UV example with 80% as the acceptance level $\pi_{\min}$. A 90% confidence level was chosen to calculate the maximum likelihood and Mee confidence interval lower bounds and $\gamma$ was fixed at 0.5 in $\beta-\gamma$ tolerance intervals.
Table 2: Numerical results of the application of the quality-level and tolerance-interval validation approaches to the HPLC-UV example.

<table>
<thead>
<tr>
<th>Concentration</th>
<th>( \mu_T )</th>
<th>( \hat{\pi} )</th>
<th>( \pi_{ML}^T )</th>
<th>( \pi_{Mee}^T )</th>
<th>( \beta_L ) low</th>
<th>( \beta_L ) high</th>
<th>( \beta_C ) low</th>
<th>( \beta_C ) high</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.35</td>
<td>0.9818</td>
<td>0.9378</td>
<td>0.8716</td>
<td>-2.55</td>
<td>2.50</td>
<td>-2.53</td>
<td>2.49</td>
</tr>
<tr>
<td>2</td>
<td>48.24</td>
<td>0.9872</td>
<td>0.9558</td>
<td>0.9053</td>
<td>-5.55</td>
<td>1.77</td>
<td>-5.56</td>
<td>1.78</td>
</tr>
<tr>
<td>3</td>
<td>437.82</td>
<td>0.9863</td>
<td>0.9428</td>
<td>0.8545</td>
<td>-54.70</td>
<td>20.97</td>
<td>-53.29</td>
<td>19.57</td>
</tr>
<tr>
<td>4</td>
<td>838.65</td>
<td>0.9841</td>
<td>0.9262</td>
<td>0.7986</td>
<td>-82.35</td>
<td>106.57</td>
<td>-74.84</td>
<td>99.05</td>
</tr>
</tbody>
</table>

Figure 4: Quality-level (left) and accuracy profiles (right) for the HPLC-UV example.

The four methods confirm the validity of the analytical method. In the quality-level approaches, the two “quality-level profiles” corresponding to the maximum likelihood and Mee approaches are above the acceptance level \( \pi_{\text{Min}} = 0.8 \). The Mee approach looks, however, much more conservative than the maximum likelihood one. The “accuracy profiles” corresponding to the two tolerance intervals types give very similar results (due to the fact that \( \gamma = 0.5 \) was chosen) and confirm the validity of the method as they are well within the +/-15% acceptance limits.

4. SIMULATION STUDY

This section compares the four validation procedures on the basis of simulations when the acceptance limit \( \lambda \) is fixed at 0.15, (i.e. [-15%, +15%]), the acceptance level \( \pi_{\text{min}} \) at 80%, the confidence level for the ML one-tailed confidence interval for \( \pi \) at 90% and \( \gamma \) at 0.5. Four valid (\( \pi=0.85 \) and \( \pi=0.95 \)) and two non valid (\( \pi=0.75 \)) hypothetical measurement processes or analytical methods were chosen, as shown in Table 3 and Figure 5. In these, three are unbiased and three are biased. The data were generated according to an ANOVA I random model with \( R=\sigma^2_\mu/\sigma^2_\nu \) equal to 1/4 and 4 for each possible value of the total variance \( \sigma^2_\nu \).

For each of the resulting 12 possible scenarios, 1000 samples of size N, with N ranging from 6 to 64, were randomly generated. Simulation results are given in Figure 6 as the proportion of cases where the method was accepted as valid with respect to sample size N. Note that, in an ANOVA I sampling scheme, N is indeed the \( n_\mu \times n_\nu \) product taken here at the following values: 2×3, 3×3, 3×4, 4×4, 4×5, 5×5, 5×6, 6×6, 6×7, 7×7, 7×8 and 8×8.
Table 3: Six scenarios used to compare the performances of the four validation procedures

<table>
<thead>
<tr>
<th>δ</th>
<th>π</th>
<th>σ_X</th>
<th>Valid/Not valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.95</td>
<td>0.0765 Valid</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.85</td>
<td>0.104 Valid</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.75</td>
<td>0.13 Not valid</td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>0.95</td>
<td>0.0605 Valid</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.85</td>
<td>0.091 Valid</td>
</tr>
<tr>
<td>6</td>
<td>0.05</td>
<td>0.75</td>
<td>0.12 Not valid</td>
</tr>
</tbody>
</table>

Figure 5: Six scenarios used to compare the performances of the four validation procedures

Figure 6: Simulation results expressed as the probability of accepting validation with respect to the total sample size for the twelve simulation scenarios and the four validation approaches (Δ (red) = ML quality level, × (blue) = Mee quality level, o (black) = θ-expectation tolerance interval and + (green) = θ−γ content tolerance interval).
The following conclusions can be drawn from these simulations:

- As expected, in most cases, the probability of accepting a valid method tends to 1 as \( N \) increases and the probability of accepting an invalid method tends to 0 as \( N \) increases. The client risk (probability of accepting an invalid method as valid) can however be quite large when \( R \) is large and \( N \) small.

- The lower bound of the confidence interval obtained for the quality level with the Mee approach is always very conservative and usually leads to lower probabilities of accepting the validity of the method than the other approaches. This approach is thus unacceptable for the laboratory, which would too often see its method rejected when it is actually valid.

- When \( R \), the ratio of the between-run and within-run variances, decreases, all the validation approaches become more efficient as they discriminate better between valid and invalid methods. This is expected since, as \( R \) decreases, the effective sample increases and a smaller number of runs is necessary to estimate the total variance \( \sigma^2 \) precisely.

- The \( \beta \)-content and \( \beta-\gamma \) content tolerance interval approaches behave very similarly when \( R \) is low, but the \( \beta-\gamma \) content approach is more conservative when \( R \) is higher, a situation that should nevertheless be avoided in practice. This must be due to the way the uncertainty on \( \sigma^2 \) is taken into account in the interval. This would deserve further investigation.

- The maximum likelihood quality-level approach behaves very well for low \( R \) as it shows lower overall client risks in the biased and unbiased cases, and better efficiency in declaring a valid method to be valid. This is less clear when \( R \) decreases, but its performance remains good overall, compared to the other approaches.

- The tolerance interval approaches suffer from an undesirable behaviour for biased methods. This can be seen for Situation 5 and \( R=1/4 \). The probability of accepting the method decreases to 0 as \( N \) increases even though, in Situation 5, the method is in fact valid. This is a “border effect” that has already been mentioned by Castaneda-Mendez [13] and discussed in detail by Boulanger et al. [4]. The latter paper shows that the acceptance region of the tolerance-interval decision rule does not coincide with the ideal acceptance region of a validation test for large \( N \)s (see Figure 7). This arises (asymptotically) for the points in the acceptance region where

\[
\lambda - |\delta| < \sigma_x \frac{z_{\frac{1 - \alpha_{\beta-mn}}{2}}}{1 - \alpha_{\beta-mn}}.
\]

This problem is however negligible for usual sample sizes where the uncertainty on the parameter estimates come to the fore.

**Figure 7:** The acceptance region where the measurement method is (asymptotically) wrongly rejected with the tolerance-interval method, while it is valid and accepted with the ML method. The six simulated scenarios are marked.

This figure also shows that, asymptotically, the shape of the acceptance region of the tolerance-interval-based acceptance rule is close to that based on the total-error concept. However the
advantage of tolerance intervals is that they take into account the value of the acceptance level $\pi_{\text{min}}$ and of the uncertainty of estimation of the bias and precision, and so control both the client and the laboratory risks better.

5. CONCLUSION AND FURTHER DEVELOPMENTS

This paper presents and compares four different approaches to assess the validity of an analytical method at a pre-study level. They are presented as alternative to classical approaches presented in the conventional literature which are more focussed on the properties of the method than on the expected accuracy of the measurement results.

These approaches are not all new but are presented here in a general variance components modelling framework and with common notations. Simulations show that they all behave quite well except the Mee quality-level estimation approach which is too conservative and not economically acceptable for the laboratory.

In a concern of simplicity and shortness, this paper does not discuss the case of unbalanced designs which can be encountered when some data are missing in the validation experiment results. Hoofman and Kringle [3] discuss this problem for $\beta$-$\gamma$ content tolerance intervals but their results would still merit to be further developed and generalized to other validation approaches. Also, all of these procedures are based on asymptotic statistical results when, in practice, laboratories usually use validation experiments with very small sample sizes. Simulations of Section 4 show, as expected, that client and laboratory risks rapidly increase for $N<20$. A closer look to the behaviour of the validation approaches in very small sample situations should highlight more precisely their properties. These two problems and other related extensions are discussed in Ourliac [14].

ACKNOWLEDGMENTS

The authors would like to thank Catherine Hammer and Anne Ourliac for their contributions to this research in the context of their master’s theses. The first author gratefully acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy).

REFERENCES


