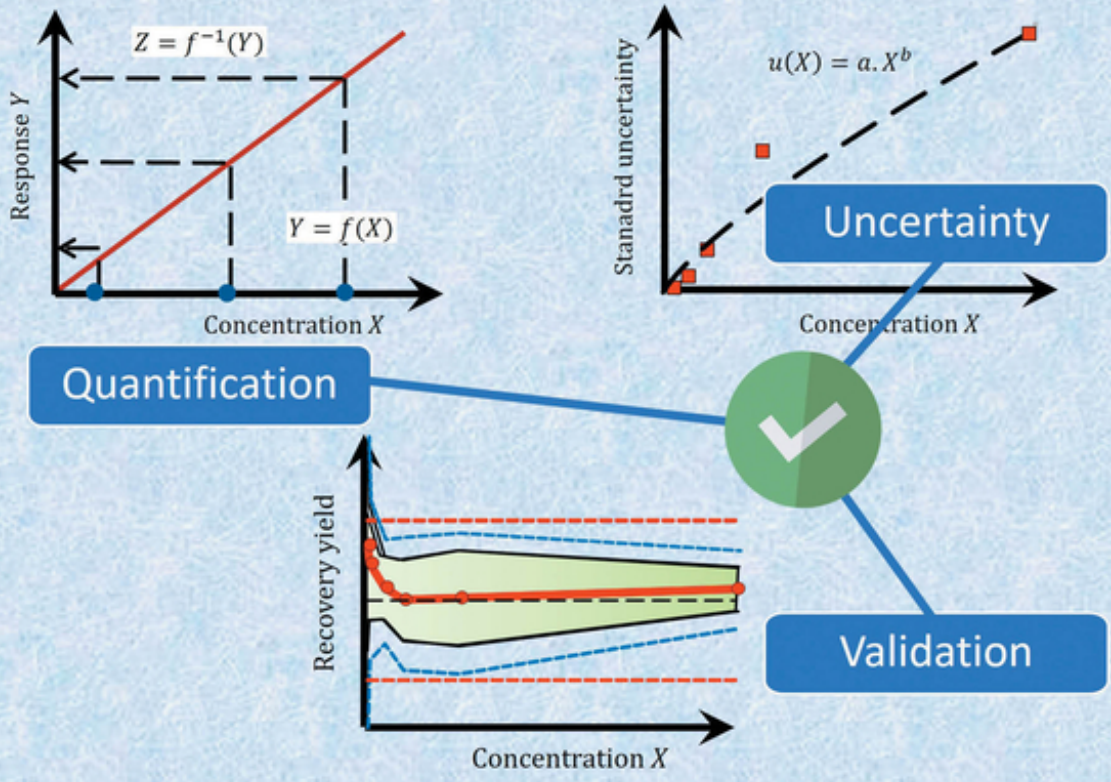
WILEY VCH

Max Feinberg and Serge Rudaz

# Quantification, Validation and Uncertainty in Analytical Sciences

## An Analyst's Companion



**Quantification, Validation and Uncertainty in Analytical Sciences**

# Quantification, Validation and Uncertainty in Analytical Sciences

An Analyst's Companion

*Max Feinberg*
*Serge Rudaz*

WILEY VCH

**Authors**

***Dr. Max Feinberg***
Paris
France

***Prof. Serge Rudaz***
Section des sciences pharmaceutiques
Université de Genève
Rue Michel Servet 1
1211 Genève
Switzerland

**Cover Image:** © Max Feinberg



**Library of Congress Card No.:** applied for

**British Library Cataloguing-in-Publication Data**
A catalogue record for this book is available from the British Library.

**Bibliographic information published by the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.



**Print ISBN:** 978-3-527-35332-3
**ePDF ISBN:** 978-3-527-84525-5
**ePub ISBN:** 978-3-527-84526-2
**oBook ISBN:** 978-3-527-84527-9

**Typesetting** Straive, Chennai, India

# Contents

**List of Figures** *xi*
**List of Resources** *xv*
**Preface** *xvii*
**Glossary of Symbols** *xxi*
**Acknowledgments** *xxiii*

**1 Quantification** *1*
1.1 Define the Measurand (Analyte) *1*
1.1.1 Quantification and Calibration *2*
1.1.2 Authentic *versus* Surrogate *3*
1.1.3 Signal Pretreatment and Normalization *6*
1.2 Calibration Modes *9*
1.3 External Calibration (EC) *10*
1.3.1 Authentic Analyte in Authentic Matrix: MMEC *10*
1.3.2 Authentic Analyte in Surrogate Matrix *12*
1.3.3 Surrogate Calibrant in Authentic Matrix *13*
1.3.4 Surrogate Calibrant in Surrogate Matrix *14*
1.4 In-sample Calibration (ISC) *15*
1.4.1 Authentic Analyte: Standard Addition Method *15*
1.5 Some New Quantification Techniques *17*
1.5.1 Isotopic Pattern Deconvolution (IPD) *18*
1.5.2 Direct Internal Calibration with Labeled Calibrant (IC-SIL) *20*
References *23*

**2 Calibration** *25*
2.1 Direct and Inverse Calibration *25*
2.2 Least-squares Regression Method *28*
2.2.1 Straight-line Computation *28*
2.2.2 Assumptions and Complements *32*
2.3 Software Implementation *34*
2.3.1 Ordinary Least-squares (OLS) Regression *34*
2.3.2 Weighted Least-squares (WLS) Regression *37*
2.4 Calibration: Special Topics *41*
2.4.1 Nonlinear Calibration Curve *41*

2.4.2 Misuses of Regression for Calibration *45*
2.4.2.1 Coefficients of Correlation and Determination *45*
2.4.2.2 Definitions of Linearity *47*
2.4.3 Statistical Aspects of Standard Addition Method (SAM) *48*
2.5 Metrological Approach to Calibration *51*
2.5.1 Errors in Inverse-predicted Values *53*
2.5.2 Calibration as a Source of Uncertainty *55*
References *57*

**3 Precision** *59*
3.1 Outputs of Interlaboratory Studies *59*
3.1.1 Diverse Precision Parameters *59*
3.1.2 Role of Series for Data Collection *60*
3.2 Analysis of Variance (ANOVA) *64*
3.2.1 Computation of Precision Parameters *64*
3.2.2 Additional Parameters *69*
3.2.2.1 Relative Standard Deviation of Parameters *69*
3.2.2.2 Variance of the Grand Mean *70*
3.3 Balanced and Unbalanced Experimental Design *71*
3.4 Software Implementation *72*
3.4.1 ANOVA Classic Algorithm *72*
3.4.2 Detect Outliers and Stragglers *75*
3.4.2.1 Other Algorithms *78*
References *79*

**4 Trueness** *81*
4.1 Trueness and True Value *81*
4.1.1 Bias and Recovery Yield *82*
4.1.2 Evolution of the Concept of True Value *83*
4.1.3 Specificity and Sources of Bias *84*
4.2 Assessment of Trueness *86*
4.2.1 Primary Operating Procedures *86*
4.2.2 Reference Materials *87*
4.2.2.1 Certified Reference Materials (CRM) *87*
4.2.2.2 External Reference Materials (ERM) *87*
4.2.2.3 Internal Reference Materials (IRM) *87*
4.2.2.4 Verification Standard Solutions *87*
4.2.2.5 Standard Addition Method (SAM) and Surrogate Samples *88*
4.3 Proficiency Testing *89*
4.3.1 Interlaboratory Comparison or Proficiency Testing Scheme (PTS) *89*
4.3.2 Organization of Proficiency Testing Schemes *90*
4.3.3 Reference Value of the Test Material *91*
4.3.4 Performance Scores *93*
4.3.5 Algorithm A *94*
4.3.6 Check Material Homogeneity or Stability *97*

4.4 Control Charts *99*
4.4.1 First Phase Assessment of the Reference Value *100*
4.4.2 Second Phase Routine Use *101*
References *102*

**5 Method Validation** *105*
5.1 Review of Validation Procedures *105*
5.1.1 Inconsistencies of Validation Vocabulary *107*
5.1.2 Validation Plans *109*
5.2 Method Accuracy Profile (MAP) *113*
5.2.1 Principles *113*
5.2.2 Method Accuracy Profile by Example *116*
5.3 Statistical Dispersion Intervals *122*
5.3.1 β-Expectation Tolerance Interval (β-ETI) *125*
5.3.2 β-γ Content Tolerance Interval (β-γ-CTI) *128*
5.4 Accuracy Profile: Special Topics *131*
5.4.1 Choose the Best Calibration Model *132*
5.4.2 Apply Consistent Experimental Design *133*
5.4.3 Check the Number of Efficient Measurements *135*
5.4.4 Select Probability Values *140*
5.4.5 Select the Type of Tolerance Interval *143*
5.4.6 Proportion of Nonacceptable Measures *144*
References *146*

**6 Measurement Uncertainty (MU)** *149*
6.1 Principle of Measurement Uncertainty *149*
6.2 General Procedure to Estimating MU *150*
6.3 Traceability at the International System of Units *152*
6.4 Stage 1. Specify the Measurand *154*
6.5 Stage 2. Identify Uncertainty Components *157*
6.6 Stage 3. Quantify Uncertainty Sources *158*
6.6.1 Type A Approach *159*
6.6.1.1 Accuracy Profile *159*
6.6.1.2 Interlaboratory Study *159*
6.6.1.3 Control Chart *160*
6.6.1.4 Proficiency Testing *160*
6.6.2 Type B Approach *160*
6.7 Stage 4. Calculate Combined Uncertainty *161*
6.7.1 Law of Propagation of Uncertainty *161*
6.7.1.1 The Model Only Contains Additions and Subtractions *162*
6.7.1.2 The Model Only Contains Products and Quotients *163*
6.7.1.3 The Model is a Complex Combination of Input Quantities *163*
6.7.2 Kragten Iterative Algorithm *165*
6.8 Calculate Expanded Uncertainty *169*
6.9 Round the Result *170*

6.10 Accuracy, Total Error, and Uncertainty *171*
6.11 Insights on Probability *174*
References *177*

**7 Measurement Uncertainty in Analytical Sciences** *179*
7.1 Published Procedures: An Evaluation *179*
7.2 Use Method Accuracy Profile Data *181*
7.2.1 Stage 1. Generic Measurement Model *181*
7.2.2 Stage 2. Generic Cause-to-Effect Diagram *182*
7.2.3 Main Sources of Uncertainty in the Laboratory *185*
7.2.3.1 Manpower *185*
7.2.3.2 Material and Handling of Items *185*
7.2.3.3 Method *185*
7.2.3.4 Machine/Equipment *186*
7.2.3.5 Environment *186*
7.2.3.6 Measurement and Other Sources *186*
7.2.4 Stages 3 and 4. Calculation of Combined Uncertainty *187*
7.3 Use Control Charts Data *191*
7.3.1 Principles of the Shewhart Control Chart *191*
7.3.2 Statistical Dispersion Intervals and Control Charts *195*
7.3.3 Estimation of the Reference Value Uncertainty *198*
7.4 Use Interlaboratory Comparison Data *200*
7.4.1 Proficiency Testing Scheme (PTS) *200*
7.4.2 Interlaboratory Studies *201*
7.5 Uncertainty Functions *203*
7.5.1 Horwitz's Model *203*
7.5.2 Fitting the Uncertainty Function *208*
7.5.2.1 How to Interpret a Power Function? *211*
7.6 Concept of Coverage Interval *211*
7.6.1 Origin of Coverage Interval *211*
7.6.2 Coverage Interval of Given Concentration *215*
7.6.3 Coverage Interval of Given Relative Uncertainty *215*
7.6.4 Obtain the Limits of the Coverage Interval *216*
References *217*

**8 Measurement Uncertainty and Decision** *221*
8.1 Framework for Decision-Making *221*
8.1.1 Decision *versus* Uncertainty *221*
8.1.2 Specification Limits and Reference Values *223*
8.1.3 Role of the Analytical Report *226*
8.2 Sample Conformity Assessment *227*
8.2.1 Define the Decision Rule *227*
8.2.2 Guard Band Concept *229*

8.3 Sampling Uncertainty *232*
8.3.1 Sampling and Heterogeneity *232*
8.3.2 Procedure of Homogeneity Check *236*
8.3.3 Example of Copper in Wheat Flour *237*
8.4 Measurement Uncertainty: Special Issues *240*
8.4.1 Influence of the Calibration Model *240*
8.4.2 Uncertainty of Corrected Results *243*
8.4.3 Increase the Number of Replicates *250*
8.4.4 Replication under Repeatability Condition *251*
8.4.5 Replication under Intermediate Precision Condition *253*
References *254*

**9 MU and Quantification Limits** *257*
9.1 Definitions and Assessment of LOQ *258*
9.1.1 Multiple Blank Standard Deviations *258*
9.1.2 Visual Examination *259*
9.1.3 Signal-to-Noise Ratio *259*
9.1.4 Empirical Experimental Approach *260*
9.2 LOQ as an Expected Relative Uncertainty *260*
9.3 Decision Limit and Detection Capability *262*
9.3.1 Concepts and Definitions *262*
9.3.2 Initial Procedure (2002) *265*
9.3.3 Modified Procedure (2021) *265*
9.3.4 Example of Calculation *267*
References *269*

**10 Examples of MU Application** *271*
10.1 Standard Addition Method and Drug Quality *271*
10.1.1 SAM Without Replication *273*
10.1.2 SAM with Replication *279*
10.1.3 Estimation from Method Accuracy Profile *281*
10.2 Method Comparison Using Uncertainty *283*
10.2.1 Analyte Defined by the Operating Procedure *283*
10.2.2 Kjeldahl and Dumas Method Comparison *285*
References *287*

**11 Conclusions** *289*
11.1 Role of the Number of Replicates *290*
11.2 Traceability to International Units *290*
11.3 Education about Uncertainty *292*
11.4 Risk Analysis *292*
11.5 Harmonization of MU Estimation Procedures *293*
References *294*

**Annexes** *295*
The 10-step MAP Procedure *295*
Glossary of Used Terms *295*
Acronyms *302*
Reference *304*

**Index** *305*

## List of Figures

**Figure 1** How to read this book. *xviii*

**Figure 1.1** Schematic representation of the quantification principle. *2*

**Figure 1.2** Schematic representation of absolute, semi, and relative quantification modes. *6*

**Figure 1.3** Contribution to the reproducibility of two quantification methods in liquid chromatography of saccharides. *9*

**Figure 1.4** Two-run standard addition method. *16*

**Figure 1.5** Calibration modes in analytical sciences. *17*

**Figure 1.6** LC-MS on endogenous metabolites: proposed workflow for selecting a calibration operating procedure. *22*

**Figure 2.1** THEOPHYLLINE – illustration of the calibration data of series 1. *27*

**Figure 2.2** Direct calibration and inverse calibration. *28*

**Figure 2.3** Principles of ordinary least-squares (OLS) method. *30*

**Figure 2.4** ELISA – determination of interleukin 6. *43*

**Figure 2.5** SAM – multiple point standard addition method. *48*

**Figure 3.1** Graphical representation of diverse total variance decomposition, affording diverse sources of variation. *60*

**Figure 3.2** LEAD – illustration of interlaboratory study. *63*

**Figure 3.3** Geometric interpretation of the general ANOVA. *66*

**Figure 3.4** (a) ANOVA – observed model. (b) ANOVA – theoretical model. *68*

**Figure 3.5** Graphical representation of the experimental design. (a) Balanced and (b) unbalanced. *71*

**Figure 3.6** Diverse types of outliers in an interlaboratory study. *75*

**Figure 3.7** LEAD – interlaboratory study after outlier deletion. *77*

**Figure 4.1** Example of systematic error generated by the integration mode of poorly resolved chromatographic peaks. *84*

**Figure 4.2** Geometric interpretation of additive and multiplicative bias. *85*

**Figure 4.3** ALFALFA *97*

**Figure 4.4** NITROGEN – examples of anomalies in a wheat flour control chart. *100*

**Figure 4.5** Possible location of different Qcontrol charts in a routine laboratory. *102*

**Figure 5.1** Frequency of terms used in validation guides. *109*

**Figure 5.2** Example of a multicriteria validation procedure. *111*

**Figure 5.3** Example of single-criterion validation procedure. *112*

**Figure 5.4** Number of publications with "accuracy profile" in the title ratioed to all "validation" published papers. *115*

**Figure 5.5** THEOPHYLLINE – MAP with six validation materials. Inverse-predicted concentrations are obtained with WLS quadratic model. *119*

**Figure 5.6** THEOPHYLLINE – validation and validated ranges ($\beta\% = 80\%$). *120*

**Figure 5.7** THEOPHYLLINE – lower part of the MAP expressed as absolute inverse-predicted concentration ($\beta\% = 80\%$). *121*

**Figure 5.8** Schematic representation for LOQ calculation. *122*

**Figure 5.9** Comparison tolerance and confidence intervals calculated for 18 replicates assumed to be normally distributed. *124*

**Figure 5.10** THEOPHYLLINE – accuracy profile with the two tolerance intervals ($\beta\% = 80\%$, $\gamma\% = 95\%$). Inverse-predicted concentrations are obtained with WLS quadratic models. *131*

**Figure 5.11** THEOPHYLLINE – accuracy profile (MAP) with the two tolerance intervals obtained for OLS quadratic models ($\beta\% = 80\%$, $\gamma\% = 95\%$). *133*

**Figure 5.12** Schematic representation of the experimental design to be used to build a relevant accuracy profile. *134*

**Figure 5.13** Influence of the experimental design parameters on the number of effective measurements. *136*

**Figure 5.14** Influence of the experimental design on the coverage factor with $\beta\% = 80\%$. *138*

**Figure 5.15** Coverage factor as a function of the number of efficient measurements. *140*

**Figure 5.16** THEOPHYLLINE – half tolerance intervals for different probability values. *143*

**Figure 6.1** The 4-step GUM general procedure for measurement uncertainty (MU) estimation. *151*

**Figure 6.2** LEAD – cause to effect diagram of the sources of uncertainty when determining lead by ICP-ID-MS. *157*

**Figure 6.3** Triangular distribution law applied to a digitized reading rounded to 90. *161*

**Figure 6.4** LEAD – uncertainty budget. *168*

**Figure 6.5** Schematic representation of the main concepts used for method validation. *172*

**Figure 6.6** Comparison between the total analytical error (*TAE*) model and the measurement uncertainty (MU) model. *173*

**Figure 6.7** Modeling: moving from the real world to idealized world. *175*

**Figure 6.8** CORTISOL – accuracy profiles from four validation studies. *176*

**Figure 6.9** CORTISOL – uncertainty functions of four accuracy profiles. *177*

**Figure 7.1** Different MU estimation procedures proposed for the analytical sciences. *180*

**Figure 7.2** Generic cause-to-effect diagram with eight main classic sources of uncertainty. *184*

**Figure 7.3** THEOPHYLLINE – 95% coverage intervals. *192*

**Figure 7.4** ALBUMIN – control chart and QC. *194*

**Figure 7.5** (a) LEAD – laboratory coverage intervals including outliers, (b) LEAD – individual coverage intervals after removing outliers. *203*

**Figure 7.6** Horwitz (solid line) and Thompson (dashed line) models. *205*

**Figure 7.7** Precision and trueness acceptance criteria proposed in various official guidelines. *208*

**Figure 7.8** (a) THEOPHYLLINE – standard uncertainty function, (b) THEOPHYLLINE – relative uncertainty function. *210*

**Figure 7.9** Different power functions when power coefficient $b$ varies, and $a = 0.2$ remains constant. *212*

**Figure 7.10** Coverage interval and measurement uncertainty. *214*

**Figure 8.1** Total circulating testosterone reference values for normal and pathological states with associated MU. *224*

**Figure 8.2** Three ways to assess sample conformity to a unilateral or bilateral specification interval. *228*

**Figure 8.3** The guard band concept introduced by the JCGM. *229*

**Figure 8.4** Influence of MU on acceptability or rejection intervals. *230*

**Figure 8.5** Basic sampling vocabulary. *232*

**Figure 8.6** Main types of spatial distribution of an analyte in a batch or population. *233*

**Figure 8.7** Cause to effect diagram of the sampling operation. *235*

**Figure 8.8** COPPER – distribution of measurements and control number. *239*

**Figure 8.9** PARACETAMOL – method accuracy profiles using two calibration models on the same data. *241*

**Figure 8.10** PARACETAMOL – uncertainty functions for the two calibration models. *242*

**Figure 8.11** THEOPHYLLINE – comparison of uncertainty functions obtained with two calibration models. *243*

**Figure 8.12** NICOTINIC – accuracy profiles before and after correction. *247*

**Figure 8.13** (a) NICOTINIC – relationship between the average correction factor and the concentration. (b) NICOTINIC –Relative uncertainty function before and after correction. *249*

**Figure 8.14** Four possible replicate definitions according to the sample preparation starting step among the sample preparation steps. *250*

**Figure 9.1** Plausible classic and alternative approaches to estimate LOD and LOQ. *262*

**Figure 9.2** Definitions of decision limit and detection capability according EU regulation [7] for a substance with a maximum residue limit (*MRL*) of 100 µg/kg. *264*

**Figure 9.3** Detection capacity for the calibration curve method according to ISO 11843-2. *269*

**Figure 10.1** Example of assay obtained by SAM on a pharmaceutical product containing TDF. *274*

**Figure 10.2** Determination of FTC, by standard additions to a pharmaceutical product announced at 200 mg/tablet. *278*

**Figure 10.3** Average levels found (mg/tablet) with the coverage intervals for six medicines noted from A to F. *278*

**Figure 10.4** MU estimates of six drug lots for two nominal strengths of 200 and 245 mg/tablet. *280*

**Figure 10.5** Accuracy profile of Dumas's method applied to dairy products using Kjeldahl method as reference. *284*

**Figure 10.6** Comparison of the uncertainty functions of both methods used to determine total nitrogen in foods. *286*

**Figure 10.7** Accuracy profile of Dumas's method applied to dairy products compared to improved Kjeldahl method. *286*

**Figure 11.1** Relationship between risk consequences and measurement uncertainty. *293*

## List of Resources

**Resource A** Linear and quadratic calibration (Excel). *35*
**Resource B** Calibration using OLS and WLS (Python). *39*
**Resource C** Nonlinear calibration (Python). *44*
**Resource D** Standard addition method (Excel). *50*
**Resource E** Precision parameters for a balanced design (Excel). *74*
**Resource F** Precision parameters for an unbalanced design (Excel). *78*
**Resource G** Algorithm A (Python). *95*
**Resource H** β-Expectation tolerance interval (Excel). *128*
**Resource I** β-γ content tolerance interval (Excel). *130*
**Resource J** Probability of non-acceptable measurements (Excel). *145*
**Resource K** Iterative algorithm applied to LEAD (Python). *165*
**Resource L** Rounding a result (Excel). *171*
**Resource M** Calculation of the coefficients of a power function (Excel). *209*
**Resource N** Coverage interval for a given concentration (Excel). *215*
**Resource O** Coverage interval for given relative uncertainty (Excel). *216*
**Resource P** Decision limit – calibration curve procedure of ISO 11843-2. *268*
**Resource Q** Calculation of the SAM extrapolated concentration (Excel). *276*

# Preface

Why an Analyst's Companion? Millions of analyses are carried out every day in laboratories for all sectors of industry and science. Many people are willing to pay for these analyses because they are considered effective in making a scientifically sound decision. Though few publications address the economics of analytical sciences, nonetheless, a report by the European Commission concluded in 2002 that "for every euro devoted to measurement activity, nearly three euros are generated" [1]. But is it easy and simple to use an analytical result, and does it always allow you to make the right decision? Some questions illustrate the risks involved in relying on a result:

- How do you know that the laboratory used the method that gave the exact result?
- Like any measurement, analysis is subject to error. How can you estimate them?
- How can a spurious measurement be used effectively?

This is the right time to explain why and how the concept of measurement uncertainty (MU) can be used to better manage these risks. This also means that a new challenge for analysts is to develop an appropriate method for estimating MU more explicitly applicable to analytical sciences. In this perspective, a tool based on the statistical dispersion intervals called method accuracy profile (MAP) is proposed as the backbone of the book. The theoretical aspects of the MAP procedure and MU estimation are presented in several examples and template worksheets to help analysts quickly grasp this tool.

At the turn of the 1970s, three analytical chemists, Bruce Kowalski, Luc Massart and Svante Wold, conceptualized a discipline they called Chemometrics [2]. Unfortunately, they all have passed away since, but their work is still vivid. Many chemometrics books have been published, proving the added value of statistics to analytical sciences. Some are globally addressing chemometrics [3–5] other are more focused on statistics [6, 7], and others on method validation [8, 9].

This book contributes to the application of chemometrics, but the obvious aim is not to repeat what is available in many valuable publications. Only a few books precisely address measurement uncertainty in analytical sciences [10–12]. They present limited facets and do not propose a more comprehensive approach. The aim of this book is to describe a global procedure for MU estimation, easily applicable in analytical laboratories. In a recent publication, we have exposed in a condensed manner

our view of the link between validation and measurement uncertainty [13]. This book develops more extensively and practically our viewpoint.

However, it is not satisfactory to simply propose a *modus operandi* (even if it is claimed to be universal) for estimating MU when this parameter is still new in analytical sciences and not always well identified by end-users. Therefore, several chapters are dedicated to its practical use in decision-making, demonstrating its advantages. These remarks indicate that this book is primarily intended for professional analysts, although researchers and students may find it of interest.

In order to reach this goal, the book is organized around practical responses covering three major questions daily put to analysts when they develop a new method or routinely apply it to unknown samples:

- How to quantify the analyte?
- How to validate the method?
- How to estimate the measurement uncertainty?

How does this book give answers these questions? We use as a roadmap a tool based on the application of statistical dispersion intervals called MAP. The latter was initially conceived for method validation, but it can easily be used for MU estimation. While method validation is often reduced to computing a set of disconnected parameters to be estimated, the MAP approach is more global. It consists in defining the interval where the method is able to produce a given proportion of acceptable results. This perspective is in harmony with the uncertainty approach proposed by metrologists some decades ago that consists in computing the so-called coverage interval of the result.

The chapters of the book can be read independently. This may explain some redundancies in the quoted publications. But they are structured according to a reading thread illustrated in Figure 1. The thick grey arrow is the backbone. Six main chapters are characterized as rounded angle boxes. Three of them are devoted to measurement uncertainty, as it is a key issue of the book.

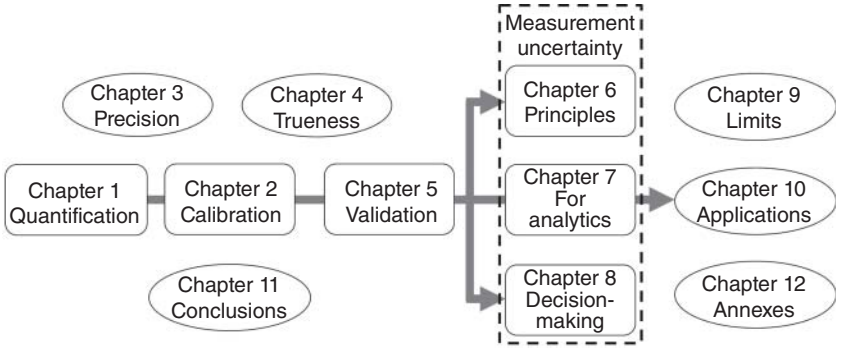



**Figure 1** How to read this book.

Additional chapters appear as ellipses. They bring two kinds of information. On the one hand, theoretical background, such as precision and trueness parameter estimation and how to compute them, may be useful to better understand statistical developments involved in the method accuracy profile. On the other hand, specific examples of MU applications. One is devoted to the limits of quantification and the challenging question of controlling samples with low analyte concentration, another to method comparison.

Several data sets provide the link between the different chapters. They are used throughout for practical data handling and real software application. The aim of this data-oriented presentation is to help the analyst apply the proposed techniques in the laboratory, in keeping with the title "Companion." This also practicality means that numerical applications for all topics covered are presented and illustrated alongside the theoretical considerations. These are based on detailed Microsoft Excel® worksheets or free equivalent, such as OpenOffice® Calc, included with the book. This software is user-friendly and does not require much explanation, and probably everyone in the laboratory knows how to use it. Although criticized by professional statisticians (for good reasons), this software is extremely helpful for quick and simple statistical computation in a laboratory, and several pitfalls can easily be avoided:

- Worksheet cell content is easily modified without any warning. Thus, once created and validated, the best initiative is to protect the worksheet or whole workbook.
- The formula inside cell is not visible unless the option to show formulas is on. To help the understanding of the template worksheets developed for this book, all formulas are made visible in the cell next to the resulting. The built-in function `FORMULATEXT` is used for this aim. It is only available in the most recent Excel releases.
- Confusion may exist between a worksheet and a text editor. Fancy presentation must be avoided, and it is better to embed a worksheet within a text editor rather than trying to do everything with a single software.

The basic use of worksheet software does not allow complex statistical calculation though it contains many built-in functions, which are used in the following examples. It is possible to use the development environment called Visual Basic for Applications coming with Excel to build more complex programs, but it requires some practice. For the most sophisticated applications, we preferred to provide Python program examples. This software is increasingly popular, and the accuracy of statistical functions is widely recognized. For instance, complex techniques, such as non-linear or weighted regression techniques, are easily implemented. Python is simpler than professional statistical software. It is developed under a free license, and there is an exceptionally large community of users who can help. The drawback is that it is a patchwork, and many additional modules must be imported to apply some methods. The simplest way to install Python is to download a free package called Anaconda [14] and select the Spyder development environment. Presented examples were programmed in this environment.

## References


1 G. Williams (2002). The assessment of the economic role of measurements and testing in modern society. *European Measurement Project*, Pembroke College, University of Oxford.

2 Wold, S. and Sjöström, M. (1998). Chemometrics, present and future success. *Chemometrics and Intelligent Laboratory Systems* 44: 3–14.

3 Kowalski, B.R. (1984). *Chemometrics: Mathematics and Statistics in Chemistry*. Dordrecht: Springer.

4 Massart, D.L. (1997). *Handbook of Chemometrics and Qualimetrics Part A*. Amsterdam: Elsevier.

5 Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C. et al. (1998). *Handbook of Chemometrics and Qualimetrics Part B*. Amsterdam: Elsevier.

6 Ellison, S.L.R., Barwick, V.J., and Farrant, T.J.D. (2009). *Practical Statistics for the Analytical Scientist a Bench Guide*, 2e. Middlesex: LGC.

7 Miller, J.N., Miller, J.C., and Miller, R.D. (2018). *Statistics and Chemometrics for Analytical Chemistry*, 6e. England: Pearson Education Limited.

8 Ermer, J. and Miller, J.H.M.B. (2006). *Method Validation in Pharmaceutical Analysis*. Weinheim: Wiley-VCH Verlag GmBH.

9 Swartz, M.E. and Krull, I.S. (2012). *Handbook of Analytical Validation*. Boca Raton, FL: CRC Press.

10 De Bièvre, P. and Günzler, H. (2013). *Measurement Uncertainty in Chemical Analysis*. Berlin, Heidelberg: Springer.

11 Bulska, E. (2018). *Metrology in Chemistry*, *Lecture Notes in Chemistry Series*, vol. 101. Springer.

12 Hrastel, N. and da Silva, R.B. (2019). *Traceability, Validation and Measurement Uncertainty in Chemistry: Vol. 3: Practical Examples*. Springer International Publishing.

13 Rudaz, S. and Feinberg, M. (2018). From method validation to result assessment: established facts and pending questions. *Trends in Analytical Chemistry* 105: 68–74.

14 Anon. (2020). Anaconda Software Distribution. Anaconda Inc. https://docs.anaconda.com/ (accessed 30 July 2023).

# Glossary of Symbols

| Symbol | Term |
|---|---|
| $\beta$ | Coverage probability of the tolerance interval |
| $u(Z)$ | Standard uncertainty $Z$ |
| $r$ | Coefficient of correlation |
| $Z$ | Inverse-predicted concentration in the working sample |
| $Z^*$ | Extrapolated sample concentration (standard addition method) |
| $Y$ | Measured instrumental response |
| $\hat{Y}$ | Predicted instrumental response |
| $X$ | Concentration of the (authentic) analyte in the working sample |
| $\overline{X}, \overline{Z}$ | Average |
| $\overline{\overline{X}}, \overline{\overline{Z}}$ | Grand average |
| $X_c$ | Concentration of the (surrogate or not) analyte in the calibrant |
| UR% | Relative expanded standard uncertainty |
| $U(Z)$ | Expanded uncertainty $Z$ |
| AIC | Akaike Information Coefficient |
| $A$ | Variance ratio $A = s_B^2/s_r^2$ |
| $\delta$ | Bias |
| $E$ | Random error variable |
| $f$ | Any calibration or uncertainty function |
| $f^{-1}$ | Inverse of any function |
| $\beta$-$\gamma$-CTI | $\beta$-$\gamma$-Content Tolerance interval |
| $\beta$-ETI | $\beta$-Expectation Tolerance interval |
| $CF$ | Correction factor |
| $p\%$ | Proportional correction factor |
| AA | Authentic analyte (used as subscript) |
| IS | Internal Standard (used as subscript) |

| Symbol | Term |
|---|---|
| $1-\alpha$ | Level of confidence (also noted $\gamma$) |
| [A−, A+] | Acceptance interval |
| $u_c(Z)$ | Combined standard uncertainty $Z$ |
| $u^2(Z)$ | Standard variance of $Z$ |
| $RF$ | Response Factor |
| $SP$ | Sum of crossed products of deviations to the mean |
| $SS$ | Sum of squared deviations to the mean |
| $s_r^2$ | Repeatability variance |
| $s_W^2$ | Within-series variance |
| $s_R^2$ | Reproducibility variance |
| $s_L^2$ | Between-laboratories variance |
| $s_{\mathrm{IP}}^2$ | Intermediate precision variance |
| $s_B^2$ | Between-series variance |
| $r^2$ | Coefficient of determination |
| $k_{\mathrm{TI}}$ | Tolerance factor of a tolerance interval |
| $k_{\mathrm{GUM}}$ | Coverage factor |
| $k_{\mathrm{GUM}}$ | Standardized coverage factor (GUM) |
| $a_0, a_1, a_2, \ldots$ | Coefficients of the calibration model |
| $[Z \pm U(Z)]$ | Coverage interval |
| $G_n$ | Input quantity of the measurement model |
| $[X_L, X_U]$ | Measuring interval or working interval |

## Acknowledgments

The authors wish to thank Professor Douglas Rutledge from the University Paris-Saclay for his careful and helpful revision of this book.

# 1

# Quantification

## 1.1 Define the Measurand (Analyte)

The initial question for the analyst is to define what is expected to be measured. According to the International Vocabulary of Metrology [1], the "quantity intended to be measured"[1] is called the measurand, or more specifically, the analyte, when considering measurement methods applied to chemical and biochemical substances. But this simple definition may be misleading while an analyte may have variable forms during the analytical process. It is not always certain that the substance finally measured is initially intended to be measured. For example, during sample preparation, the initial organic form of the analyte may change to inorganic, and what was intended to be measured is finally modified. For instance, in living organisms, heavy metal is present combined with proteins, such as mercury to metallothionein. Still, when analyzed after mineralization, it can be transformed into sulfate, perchlorate, or nitrate.

A well-known catastrophic example is the Minamata disease; when looking for mercury in food samples, the oldest methods were based on the complete sample mineralization to obtain mercury nitrate. Soon after, it was realized that the toxic forms of mercury were organic derivates. Hence, so-called total mercury had no great toxicological interest compared to the different organic forms. Speciation techniques in mineral analysis or chiral chromatographic methods are good examples of innovative approaches devoted to better maintaining the analyte in its expected form. Therefore, quantification in analytical sciences is often less straightforward than claimed. From the metrological point of view, the difficult traceability of chemical substances to international standards is one of these obstacles.

This is detailed in Section 6.3 as an introduction to the estimation of measurement uncertainty (MU) among many other sources of uncertainty. The encapsulated conception of modern and highly computerized instruments may also prevent the analyst from assessing what is measured. Digits displayed on the instrument screen represent what is "intended to be measured." The paradoxical consequence is that discussing the true nature of the analyte is often avoided, while more attention

1 Definitions or quotations extracted from standards or official documents are between double quotes.

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*, First Edition. Max Feinberg and Serge Rudaz.

should be paid to this question. The goal of this chapter is to propose things to consider on this topic. Many examples are based on mass spectrometry (MS) hyphenated methods because several are now considered highly compliant from a metrological point of view.

### 1.1.1 Quantification and Calibration

The metrology motto could be measuring is comparing. Therefore, when quantifying an analyte, the comparison principle must be previously defined. This preliminary step is usually called calibration. In modern analytical sciences, most methods use measuring instruments ranging from simple, specific electrodes to sophisticated devices; therefore, calibration procedure may enormously vary according to the nature of the instrumentation. This chapter attempts to classify the different quantification/calibration strategies applied in analytical laboratories. Because this subject is not harmonized, the employed vocabulary may vary from one domain of analysis to another and be confusing. For each term, we tried to give a definition, but it may be incomplete due to the considerable number of analytical techniques. Many suggested definitions are listed in the glossary at the end of the book.

Whatever the measuring domain, classic differences are made between direct and indirect measurement techniques. Direct method can usually refer to a measurement standard, for instance, when measuring the weight of an object on a two-pan balance with standard weights. Indirect measurements are performed using a transducer, a "device, used in measurement, which provides an output quantity with a specified relation to the input quantity."

Reversely, with a one-pan balance, measurements are indirect. At the same time, result is obtained by means of a mathematical model linking the calibrated piezoelectrical effect on the beam to the weight. In analytical sciences, methods are usually indirect. Some exceptions are set apart, classified as direct primary operating procedures by BIPM (Section 4.2.1). For most chemical or biological analytical techniques, the measuring instrument must be calibrated with known reference items before use. Finally, quantification involves three elements, as outlined in Figure 1.1:

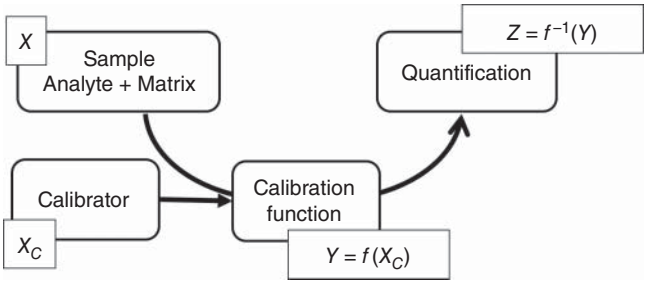



**Figure 1.1** Schematic representation of the quantification principle.

- The analyte is in the working sample. Its concentration is denoted $X$. The searched compound (chemical or biological) is embedded within the sample matrix. It is only before any treatment that the analyte is present in the intended form. The role of sample preparation is to eliminate a large part of the matrix and concentrate on the analyte. But it may change the analyte chemical form; for instance, with the speciation of organic forms of heavy metals, sample preparation is quite different from classic mineralization.
- The calibration items are also called *calibration standards* or *calibrators*. They are prepared by the analyst to contain a known amount of a calibrant as similar as possible to the analyte. To underline this difference, it is denoted $X_c$. The selection of the adequate calibrant is a key-issue of quantification extensively addressed in the rest of this chapter.
- The calibration function that links the instrumental response $Y$ to the known quantity $X_c$, denoted $Y = f(X_c)$.

Figure 1.1 is an attempt to recapitulate a generic quantification procedure. Most of the time, calibrators are artificially prepared and used to build the calibration function $f$ which generally is *inverted* when analyzing an unknown sample. The three elements may be subjected to variations. Mathematical notation underlines the dissimilar roles they play for the statistical modeling of calibration and possible relationships that link the instrumental signal to the calibrant concentration. Denoting $Z$ the predicted concentration of a sample emphasizes the role of inversing calibration function as discussed in Section 2.1. Finally, considering a given calibration dataset, distinct functions $f$ can be fitted. A principal issue will be to select the best one because it deeply affects the global method performance. The goal of the present chapter is to describe some classical or new quantification procedures.

### 1.1.2 Authentic *versus* Surrogate

To be explicit, it is convenient to define some terms. If the chemical substance sought in the sample is called *authentic*, obviously, for many methods it is possible to prepare the calibrators with the authentic analyte. But other quantification methods exist based on a different calibration compound, which will be called surrogate standard or calibrant. It would be paradoxical to call it surrogate analyte, whereas the analyte can only be authentic. Therefore, when the analyte and the calibrant are different, it is necessary for the analyst to cautiously verify if they have equivalent analytical behavior and define an eventual adjustment method, such as a correction factor.

The measuring instrument is a transducer that converts the amount or the concentration of a chemical substance into a signal – usually electrical – according to a physical or chemical principle. How quantitative analyses are achieved varies from simple color tests for detecting anions and cations through complex and expensive instrumentation for determination of trace amounts of a compound or substance in a complex matrix. Increasingly, such instrumentation is a hybrid of techniques for separation and detection that requires extensive data processing.

The subject of analytical sciences has become so wide that complete coverage, providing clear information to an interested scientist, can only be achieved in a multi-volume encyclopedia. For instance, Elsevier published in 2022 the volume n°98 of the *Comprehensive Analytical Chemistry* handbook started in the 1980s.

The major obstacle in analytical sciences is the structural or chemical differences that exist between the analyte present in the working sample and the substance used as a calibrant. The instrument signal may depend on the authentic or surrogate structure of the analyzed substance: this dependence is marked with modern instrumentation such as mass spectrometers. On the other hand, the analyte present in a working sample is embedded with other chemicals, customarily called a matrix by the analysts. It is not always possible or easy to use the sample matrix when preparing the calibrators. These remarks lead to the definitions of four different quantification elements that can be combined to prepare or selecting calibrators and consequently obtain the calibration curve:

| | |
|---|---|
| Authentic analyte | The same molecule or substance present in the working sample may be available for calibrator preparation, considering a high degree of purity. |
| Surrogate standard or calibrant | This is a reference substance that is assessed and used as a reasonable substitute for the authentic analyte. For instance, in bioanalysis, it is frequent to have metabolites or derivates of the analyte that must be quantified without the reference molecule. Labeled molecules used in many methods involving isotopic dilution have recently been considered appropriate calibrants. |
| Authentic matrix | The simplest situation for using an authentic matrix is to prepare calibrants by spiking test portions of the working sample. For some applications, such as drug control, it is also possible to prepare synthetic calibrants with the same ingredients as the products to be controlled. |
| Surrogate matrix | This medium is considered and used as a substitute for the sample matrix. For instance, bovine serum is used in place of human serum. Then, it is assumed its behavior should be similar to the authentic matrix throughout the analytical process, including sample preparation and instrumental response. |

When the surrogate matrix does not behave as the authentic or when calibration is achieved without the sample matrix, matrix effects may produce bias of trueness, as explained in Section 4.1.3. More precisely, calibration standards can be prepared with several classes of matrices. Matrix classification is widely based on analyst expertise and depending on the application domain, matrix grouping is extremely variable. For instance, broad definitions applicable to biological analysis can be as follows:

| | | |
|---|---|---|
| Authentic matrix (or real) | For biological analysts, serum, urine, saliva, or stool are different classes of matrices. In food chemistry, when determining the total protein, fatty and starchy foods are classified as different, or drinking water and surface water is different for water controllers. | |
| Surrogate matrix | Matrix used as a substitute for authentic matrix. | |
| | Neat solution | Water, reagents used for extraction or elution, etc. |
| | Artificial matrix | Pooled and homogenized samples, material prepared by weighting when the composition of the authentic matrix is fully known, etc. |
| | Stripped matrix | Specially prepared materials are free of impurities or endogenous chemicals. They are mainly used for biomedical analysis. |

It can be assumed that the combined use of surrogate standard and/or surrogate matrix may induce bias. It is necessary to cautiously verify if their analytical behavior is comparable to authentic ones. At least four combinations of the above-defined quantification elements are possible, each having pros and cons as explained later. It is possible to categorize different quantification modes depending on the selected combination:

| | |
|---|---|
| Quantitative | Calibrators are prepared with authentic analytes and an authentic matrix. The amount or concentration of the analyte may be determined and expressed as a numerical value in appropriate units. The final expression of the result can be absolute, as a single concentration value; non-absolute, as a range or above or below a threshold. |
| Semi-quantitative | Surrogate standards and matrix are used. Some authors consider semi-quantitative analyses the ones performed when reference standards or the blank matrix are not readily available. |
| Relative | Sample is analyzed before and after an alteration or compared to a control situation. The relative analyte concentration is expressed as a signal intensity fold change. It is ratioed to another sample used as a reference and expressed as a signal/concentration. |

It must be clearly stated that it is impossible to strictly separate quantification from calibration since they are interdependent. According to the nature of the calibration standard used, which can be authentic or surrogate, and the matrix, which can be authentic, surrogate, neat, etc., different quantification strategies were

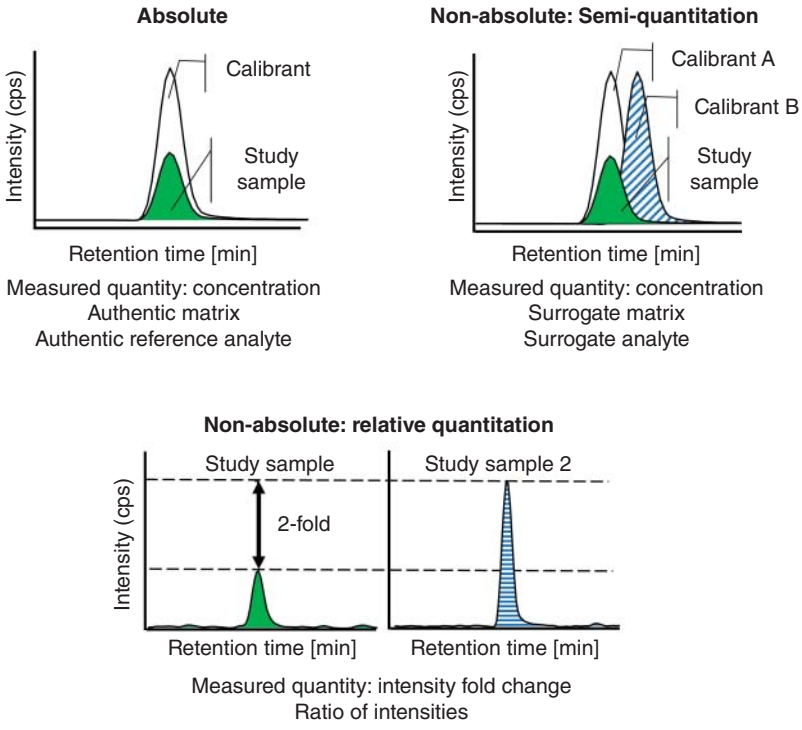



**Figure 1.2** Schematic representation of absolute, semi, and relative quantification modes.

developed to obtain the effective calibration function. A schematic overview of the differences between principal quantification modes is summarized in Figure 1.2 and more extensively explained in the rest of the chapter.

### 1.1.3 Signal Pretreatment and Normalization

Nowadays, it is quite uncommon to use the analogic electrical signal output from the measuring instrument to build a calibration model. Digitalizing signals in modern instruments opened the way to many pretreatments, such as filtering, background correction, and smoothing. It is sometimes invisible to the analyst, although this can modify the method's performance. The outcome of many methods can be complex signals such as absorption bands or peaks in spectrophotometry or elution peaks in chromatography.

This raw information is not directly used as $Y$ variable to build the calibration model; it is preprocessed. When dealing with absorption peaks, it is classic to select one or several wavelengths considered to be most informative. For instance, in biochemistry, protein concentration can be quickly estimated by measuring the UV absorbance at 280 nm; proteins show a strong peak here due to tryptophan and tyrosine residue absorbance. This can readily be converted into the protein concentration using Beer's law.

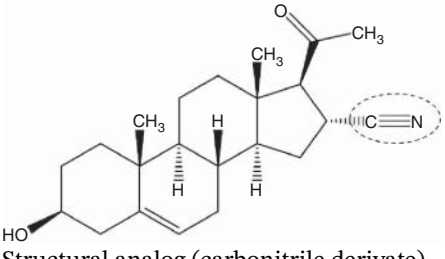
When obtaining poorly resolved absorption bands, as in near infrared spectroscopy (NIRS), the selection of one specific wavelength is difficult, and the use of a multivariate approach has been promoted. Many publications in chemometrics literature are addressing this issue. The multivariate calibration based on partial least-squares regression (PLS) has now become a routine procedure.

If the output signal is time-resolved, such as liquid or gas chromatographic peaks, they are always pretreated by an integrator. Initially, it was a separate device, but now it is included in the monitoring software. It can determine several parameters characterizing the elution peak, such as retention time at the highest point, skewness, peak height, but mainly peak area. The peak area is in the favor with analysts. But several publications demonstrated that for some methods, peak height is preferable to peak area and that when standardizing a method, the integration conditions must be carefully harmonized [2].

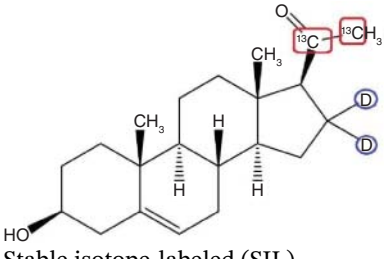
For some methods, such as MS-coupled methods, the measured response $Y$ can strongly vary according to the detector performance, such as mass analyzer type, ionization modes, ion source parameters, system contamination, ionization enhancement or suppression due to the sample matrix effect, along with other operational variables related to the analytical workflow.

Thus, the analyte relative response is standardized to compare performance over time. A common operation is adding an internal standard (IS) to the study and calibration samples at fixed concentrations. For instance, two official inspection bodies advise evaluating the matrix effects when a complex surrogate matrix is used [3, 4]. For the latter, the Food and Drug Administration (FDA) suggests investigating the matrix effect by performing parallelism testing between linear calibration curves computed with the authentic and surrogate matrices. This method is not always effective, while parallelism statistical testing is conservative, i.e. depending on the data configuration significant difference may be considered nonsignificant and only applicable to linear models.

Conversely, the European Medicines Agency (EMA) provides full instructions on how to do it and recommends comparing the extraction recovery between the spiked authentic matrix and surrogate matrix used for the calibration, along with the inclusion of IS as an easy and effective method to correct biases between these two matrices. When the analyte and the IS are affected similarly during the analytical process, instrument signals can be correctly standardized. A comprehensive approach is proposed further using the method accuracy profile (MAP); it is also an effective approach to detect and control matrix effects.



Structural analog (carbonitrile derivate)



Stable isotope-labeled (SIL)

Two main categories of IS, namely structural analogs and stable SIL, can be identified. The molecule of pregnenolone is used to exemplify this. The first category, visible on the molecule on the left, is related to compounds that generally share structural or physicochemical properties similar to the authentic analyte.

The second category, exemplified by the molecule on the right, includes stable isotopic forms of the analyte, usually by replacing hydrogen $^{1}H$, carbon $^{12}C$, or nitrogen $^{14}N$ with deuterium $^{2}H$, $^{13}C$, or $^{15}N$, respectively. Obviously, using labeled IS requires the coupling to a mass spectrometer. Deuterated IS are widely used due to their lower cost. Still, their lipophilicity increases with the number of substituted $^{2}H$, leading to differences in their chromatographic retention times with the corresponding authentic analyte. This phenomenon, known as *deuterium effect*, can also impact the instrumental response or behavior (e.g. the electrospray ionization process in MS) compared to unlabeled compounds.

Even if an increasing number of high-quality SIL are commercially available, they are limited to the most commonly used chemical compounds. When many analytes must be simultaneously quantified, the possibility of using one IS for multiple analytes should be carefully evaluated. For quantification purposes, using one IS per target compound is generally recommended when available because they are assumed to compensate for specific differences in matrix effect and extraction recovery between the calibration methodology and working samples.

To complete this rapid overview, when compatible with the analytical method, the use of standards linked to the International System of Units (SI) is a convenient means of standardizing the instrumental response and correcting the overall variation in the measurement process resulting from diverse sources of uncertainty, such as sample preparation or interfering compounds, also known as the matrix effects. The absolute instrumental response is then normalized as a response ratio:

Normalized response ratio

$$Y = \frac{Y_A}{Y_{IS}} \tag{1.1}$$

In this formula, $Y_A$ and $Y_{IS}$ are the responses obtained with the analyte and the IS, respectively. This formula gives a relative instrumental response but does not consider the respective concentrations. To be more in harmony with Figure 1.1, $Y_{IS}$ is equivalent to $Y_c$. This new notation is used because the IS is a particular example of a compound used for calibration.

The influence of signal preprocessing, such as peak integration, was experimentally demonstrated during an interlaboratory study on determining fructose, maltose, glucose, lactose, and sucrose in several foods by liquid chromatography [5]. A specific experimental design was developed to achieve this demonstration. Participants were requested to send their results calibrated as both peak heights and areas. Considering the mean values obtained with the two approaches, differences ranged from −18% up to +5%. This indicates that trueness may be affected by the quantification mode. Precision, expressed as the reproducibility variance, was computed using both sets of results.

More details about this common parameter of precision are given in Section 3.2.1. In Figure 1.3, a subset of interlaboratory results is reported. Food types are indicated

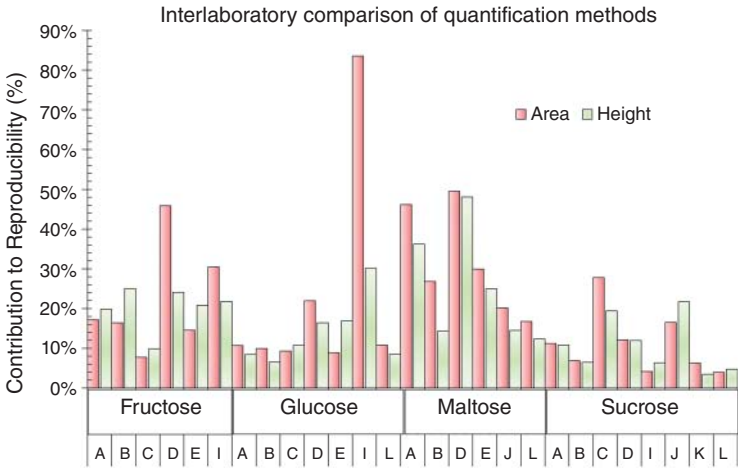



**Figure 1.3** Contribution to the reproducibility of two quantification methods in liquid chromatography of saccharides.

by an uppercase letter ranging from A to L; they are saccharide-containing processed foods, such as soft drinks, baked foods, or candies. Precision for peak area appears as vertical red bars and peak height as light green bars. The role of the signal processing method is expressed as a relative contribution to the reproducibility variance. The contributions and their differences are sometimes ridiculously small, such as fructose in food C where it is below 10%. But sometimes very impressive, such as glucose in food I. If some food is not present on the diagram, the analyte was not detected. For instance, *L* is a chocolate bar that contains no fructose. Peak area is not always the best way to quantify the analyte. In the publication, an explanation is given why the discrepancies exist. It mainly depends on the resolution of peaks and their relative values.

Detecting a peak beginning and end is a contingent subject and a source of uncertainty for the surface integration, as explained in Section 4.1.2. Finally, integrator settings can be used to optimize the integration algorithm and accordingly influence the global performance of the method.

## 1.2 Calibration Modes

Two major calibration modes are used in laboratories, namely:

| | |
|---|---|
| External calibration (EC) | A calibration curve is established independently from the working samples, whatever the calibrant nature and preparation. A single calibration function is used to quantify many samples. This is the most classical procedure, and several variants exist. |

| | |
|---|---|
| Internal calibration (IC) | The term is applied to diverse procedures. The calibration is achieved with a calibrant under different forms in the working samples. Conversely, one calibration function is obtained for each working sample to be quantified. Recently novel procedures have been developed for MS-based analysis and are detailed in Section 1.5. |

As briefly mentioned before, the analyte nature, the availability of the working sample material and the calibration material influence the selected type of calibration. This can be summarized by this simple table leading to at least four different basic configurations.

| | | Matrix | |
|---|---|---|---|
| | | **Authentic** | **Surrogate** |
| Analyte | Authentic | Yes | Yes |
| | Surrogate | Yes | Yes |

Table 1.1 attempts to classify different calibration modes, external *versus* internal, commonly used in the laboratory, including the advantages (pros) and limitations (cons) for each. As illustrated, external calibration (EC) methodologies depend on the availability of both analyte and matrix. For the procedure called in-sample calibration (ISC) there is no need to select a particular calibration matrix as the working sample matrix is used. Still remains the question of the analyte's availability. The abbreviation ISC is introduced to make the difference with internal calibration.

## 1.3 External Calibration (EC)

### 1.3.1 Authentic Analyte in Authentic Matrix: MMEC

External calibration (EC) corresponds to the most often-used operating procedure because it allows the rational determination of several routine samples with one pre-determined calibration function $Y = f(X)$. The first situation, sometimes called matrix-matched external calibration (MMEC), represents a good metrological quantification approach and is extensively discussed in the major international guidelines to validate bioanalytical methods [6].

With exogenous substances, such as rare pollutant chemicals, a blank matrix is generally available and permits EC with authentic analyte in a representative matrix. On the other hand, with endogenous compounds at endogenous concentration, such as vitamins in foods, other approaches should be explored to overcome the absence of an analyte-free matrix. In this complicated context, alternative procedures have been proposed, such as background subtraction or the use of surrogate matrices and/or analytes as described below.

**Table 1.1** Proposals for a classification of calibration procedures.

| External calibration (EC) | | | | |
|---|---|---|---|---|
| **Ref.** | **Authentic analyte** | | **Surrogate standard[a)]** | |
| **Matrix** | **Authentic** | **Surrogate** | **Authentic** | **Surrogate** |
| Method | Matrix-matched (MMEC)[b)] | Surrogate matrix | Surrogate analyte | Surrogate analyte and matrix |
| Pros | Matrix effect and selectivity close to sample. | Suitable for low concentration compounds. | LOQ Lower than the background subtraction. | When authentic analyte difficult to obtain. |
| Cons | LOQ define by endogenous concentration. | Production of analyte free matrix. Possible differences in extraction recovery and matrix effect. | Accuracy depends on surrogate specificity. Additional experiment for linearity and LOQ. | Accuracy depends on surrogate specificity.<br>High differences for recovery yield to be expected. |

| In-sample calibration (ISC) | | | |
|---|---|---|---|
| **Ref.** | **Authentic analyte** | **Surrogate standard (calibrant)** | |
| | | Partially labelled isotope analogue | Fully labelled isotope or structural analogue |
| Matrix | Authentic | Authentic | Authentic |
| Method | Standard addition method (SAM) | Isotopic pattern deconvolution (IPD) | Internal calibration (IC) |
| Pros | Same matrix effect and selectivity as the sample. | High potential for accuracy | High potential for accuracy (SIL) |
| | | Relying on isotopic distribution alteration. | Reduced numbers of calibrators. |
| Cons | Need for large initial specimen volume. | Depends on analogue concentration and stability. | Depends on analogue concentration and stability. |
| | Not easy implemented for high throughput. | Additional experiment for linearity and LOQ. | Structural analogues cannot compensate for differences in ionization.<br>Additional experiment for linearity and LOQ. |

a) Isotope labelled or structural analogue.
b) With or without background subtraction.

The use of authentic matrix for multipoint EC provides an extraction recovery yield that is close to the specimen and is commonly performed to quantify exogenous substances when a large amount of the matrix is available. In the presence of endogenous compounds, a representative pooled matrix fortified with authentic calibration standards can be prepared to estimate and remove the endogenous background signal. This approach, known as background subtraction, uses the pooled matrix-matched EC to interpolate the concentration in the working samples.

As described in Section 2.2, $Z$ is the inverse-predicted concentration. It is obtained by inversing the equation of the calibration curve. Equation (2.24) illustrates the rationale in the case of a linear calibration curve where the slope $a_1$ and intercept $a_0$ refer to the regression parameters of the added authentic standards in the pooled authentic matrix.

However, the upper limit of quantification (ULOQ) as defined by several regulatory documents may be impaired by the blank response $a_0$, because detector saturation may occur. Similarly, endogenous metabolite concentrations may vary due to intra- and inter-sample variation, leading to highly variable results when a pooled matrix is used. To overcome these drawbacks, several calibration curves using different representative pooled matrices can be prepared to select the calibration model that best covers the concentration to be analyzed. MMEC cannot always correct the matrix effect when it differs between working samples, emphasizing the importance of using an IS to correct this bias.

### 1.3.2 Authentic Analyte in Surrogate Matrix

As stated, a surrogate matrix could be used as a substitute to prepare calibrants with the authentic analyte or a mixture of analytes. It can be of various complexity. For instance, in bioanalysis, several matrices are proposed as surrogates, namely neat solutions, synthetic or stripped matrices.

- Neat solutions: it can be the mobile-phase solvent mixture, extraction reagents or pure water.
- Synthetic matrices: they are composed of salt, sugar and simulate authentic matrix properties, such as analyte solubility, extraction recovery and matrix effect. When the working sample matrix is comparable to water, saliva, urine, tears and cerebrospinal fluid, neat and artificial solutions can be used as surrogate matrix.
- Stripped matrices: they can be in-house made or commercially available, such as depleted human or bovine serum. Charcoal stripping removes nonpolar material such as lipid-related materials, mainly hormones and cytokines, leading to an analyte-free matrix that can be used as a blank for the preparation of calibrators. It is important to emphasize that charcoal depletion is nonselective and may result in approximate matrix similarity.

Whatever the chosen solution, it must be shown it has the same, or comparable, extraction properties as the authentic matrix.

Hence, surrogate matrices may not perfectly simulate the original matrix. To correct those matrix biases, a proper evaluation should be performed as recommended

by both FDA and EMA guidelines. To assess the applicability of any surrogate matrix the classic requirement is to compare the slopes of the calibration curves calculated with the surrogate matrix and authentic matrix. Diverse statistical treatments are available, such as analysis of variance.

But only EMA specifies how to assess the matrix similarity by using the concept of acceptance. This consists in ratioing the slope between authentic analyte in authentic matrix *versus* authentic analyte in surrogate matrix. The obtained value should be within ±15% of the nominal value. Example of possible procedures is fully described in Section 2.4.3 and illustrated in the worksheet named Resource D. The standard addition method (SAM) is one other dedicated tool to achieve this goal.

### 1.3.3 Surrogate Calibrant in Authentic Matrix

In situations where calibration is performed using a surrogate standard, it is assumed that the physicochemical properties of both authentic analyte and the surrogate calibrant are equivalent. For instance, the extraction recovery, the chromatographic retention behavior, and the instrument response should be either identical or have acceptable differences to be fully exploited. The choice of surrogate calibrant is essential to accurately quantify the authentic analyte.

For example, ICH guidelines [7] suggest using SIL molecule as surrogate calibrant in authentic matrix, while FDA guidelines [3] do not endorse this methodology. Because the calibration reference compound does not correspond to the authentic analyte, the ratio of responses between surrogate and analyte should be investigated over the desired dynamic range. Before routinely using the surrogate calibrant, the response factor *RF* must be evaluated as an analyte-to-calibrant ratio where $X_{AA}$ and $X_{SS}$ are the concentrations of authentic analyte and surrogate standard, respectively, and corresponding instrument responses:

Response factor (analyte *versus* surrogate)

$$RF = \frac{Y_{AA}}{X_{AA}} \times \frac{X_{SS}}{Y_{SS}} \tag{1.2}$$

To achieve the appropriate *RF* estimation, different proportions of analyte/surrogate must be investigated. For MS methods, this step is compulsory to evaluate the ionization efficiency whereas the *RF* must be constant over the method working domain. Another way to investigate the *RF* is to check if both lines are parallel. It consists in comparing the slopes of the authentic analyte line and the surrogate, both performed in the same pooled matrix.

Additionally, if the *RF* is not constant over the validation domain corrections, such as LC gradient or MS/MS transitions (de)-optimization, can be investigated to obtain a balanced response. If SIL is used as surrogate calibrant, the analyst should explore the potential presence of crosstalk interferences such as isotopic pattern overlap or impurities coming from SIL standards [8]. In MS, crosstalk occurs when ions from one scan event are still present in the collision cell when a second transition is taking place. This leads to signal artifacts in the next transition's chromatogram.

The *RF* can diverge from unit value when SILs containing enriched hydrogen atoms are used, but as long as the unit value slope remains within the ±15% acceptance interval compared to the authentic analyte slope, investigated SILs can be selected as surrogate calibrants. For example, tryptophan was successfully quantified in plasma with a relative bias between −2.0 and −8.0% using its deuterated analogue, even if the response factor was 0.67 [9]. Once the *RF* has been established, a multipoint calibration is performed in a pooled authentic matrix and the concentration of the authentic analyte is computed as follows:

Corrected concentration of authentic analyte

$$Z = \frac{\left(\frac{Y}{Y_c}\right) - a_0}{RF \times a_1} \tag{1.3}$$

where:

- $Y$ and $Y_c$ refer to the measured signal $Y$ of the authentic analyte and the IS, respectively.
- Coefficients $a_1$ and $a_0$ characterize the slope and intercept of the calibration line performed with the surrogate standard.

Likewise, MMEC's use of an IS remains strongly recommended to correct for sample preparation and matrix effect variation between working samples and calibrators, thus improving trueness and precision when dealing with routine sample determination. Because the endogenous concentration of the authentic analyte in a pooled matrix is stable, an exciting possibility to implement this quantification method is to use this signal as an IS to normalize the instrument response of the surrogate standard calibration.

This approach, called Isotope Inversion, provides the same quantitative results for steroid determination as using the authentic analyte in a surrogate matrix such as active-charcoal stripped serum in this application [10]. When no signal from endogenous analyte interferes with the surrogate signal, the surrogate calibrant in authentic matrix can be a suitable alternative to the matrix-matched external calibration, especially when high endogenous concentration is present and/or intra- and inter-sample variations are observed.

### 1.3.4 Surrogate Calibrant in Surrogate Matrix

The increased commercial availability of SILs has raised interest in their use as surrogate calibrants in surrogate matrices to reduce calibration preparation time. Numerous publications have demonstrated their benefit, especially when MS detection is considered. This semi-targeted quantitation approach could be used to determine the amount of target analytes without needing authentic chemical standards. For instance, exogenous compounds were selected as potential surrogate calibrants in several biological matrices such as blood, plasma, urine, cerebrospinal fluid, and tissue homogenate [11]. In some cases, the combination of the surrogate calibrant in surrogate matrix allows extending the number of analytes that can be quantified in a single analysis.

## 1.4 In-sample Calibration (ISC)

In contrast to EC, in-sample approach calibration (ISC) is characterized by an analytical calibration function obtained directly in each working sample. The SAM is probably the most established ISC procedure and popular in many fields, such as foods, environment, or forensic toxicology, where matrices are extremely variable, when the authentic analyte is available. Two other approaches also aim to simplify the quantification condition, depending on the chemical purity and the physicochemical proprieties of surrogate calibrant such as SIL. The former predicts the authentic analyte concentration by altering its natural isotopic pattern with a labeled analog standard. The latter is applicable when no significant interferences between the analyte and SIL are observed. In this case the authentic analyte concentration is directly determined.

### 1.4.1 Authentic Analyte: Standard Addition Method

As an operating procedure for absolute quantification, SAM consists in collecting the responses of authentic analyte additions in a series of aliquots obtained from the working sample. The simplest experimental design of SAM comprises a minimum of two runs described in Table 1.2. Notations are the same as in Figure 1.1:

- Level 0, or $X_0$, is the *no-addition* level and consists in recording the response $Y_0$ in the working sample without any addition of the authentic analyte.
- Level 1, the working sample is spiked with a known amount of the authentic analyte.

By combining the two couples of data, the corrected concentration of the working sample is given by Equation (1.4).

Corrected concentration

$$Z = Y_0 \times \frac{X_1 - X_0}{Y_1 - Y_0} \tag{1.4}$$

The short worksheet below gives an example of computation. The formula applied in cell `B5` is shown in cell `C5`. Figure 1.4 illustrates the data and shows that the corrected concentration corresponds to the extrapolation where $Y$-value is zero, and the line cuts the $X$-axis.

| | A | B | C |
|---|---|---|---|
| 1 | **Simple SAM** | | |
| 2 | Concentration *X* | Response *Y* | |
| 3 | 0 | 32 | |
| 4 | 10 | 205 | |
| 5 | Corrected result | 1.8497 | =B3*(A4-A3)/(B4-B3) |

**Table 1.2** Two-run experimental design of standard addition method.

| | **Concentration *X*** | **Response *Y*** |
|---|---|---|
| Level 0 (no addition) | $X_0 = 0$ | $Y_0$ |
| Level 1 (spiked) | $X_1$ | $Y_1$ |

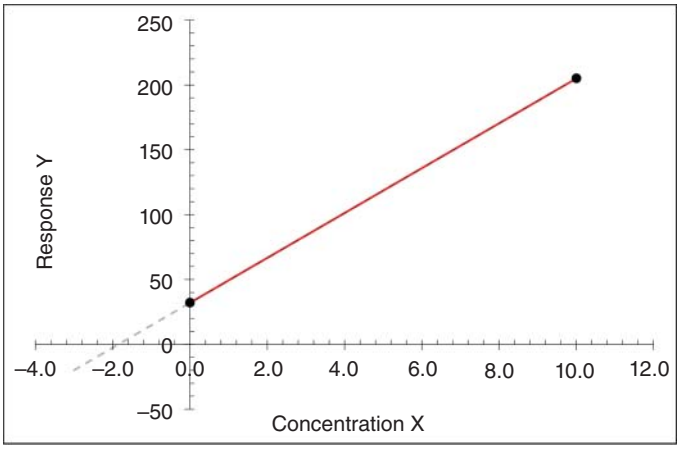



**Figure 1.4** Two-run standard addition method.

This simplified experimental design can be routinely applied when each sample may have a specific matrix effect. For instance, when analyzing surface waters, it is classic to use simple SAM for each sample because the composition is recognized as highly variable. In this case, the result is obtained by combining two measurement values that are not replicated. A discussion about the role of replication in reducing MU is presented in Section 8.4.3.

Even simplified SAM is time-consuming with preparing and measuring two test portions per working sample. The benefit is to consider interindividual differences in matrix composition, to overcome matrix effects, and avoid building an EC curve. In that respect, it can be asserted as an absolute quantification method, as far as the response is exactly proportional to the concentration, in other words, linear.

As mentioned before, the FDA suggests applying SAM in a more complex experimental design to verify if using a surrogate matrix or analyte is justified. It calculates two calibration lines: one prepared by spiking several test portions of the working sample, the other by preparing calibrators with the surrogate matrix, which can be neat.

In the classic operating procedure, the working sample is divided into four and six identical aliquots, and a fortified calibration curve is obtained by spiking increasing known amounts of the authentic analyte, e.g. 50, 100, and 200% of the expected endogenous concentration. Only the first aliquot remains nonspiked, and

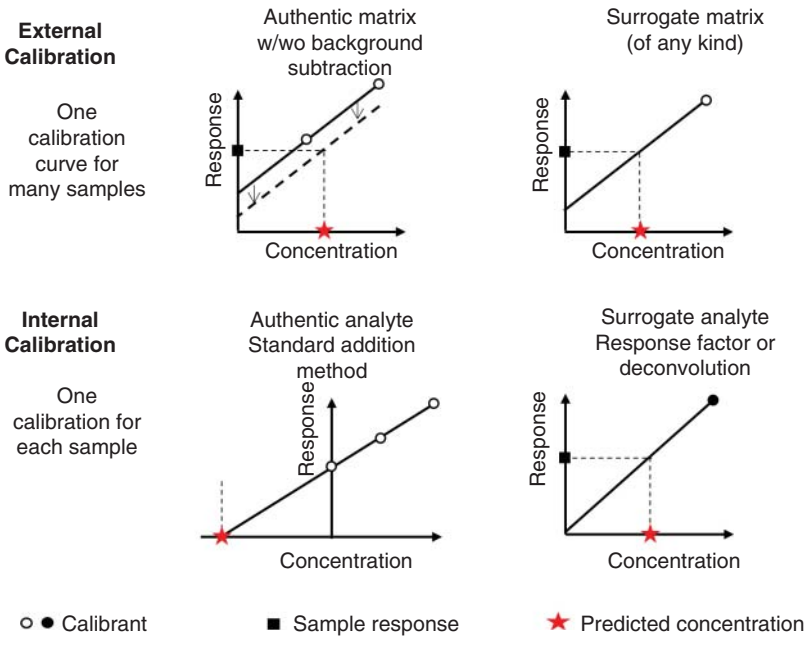



**Figure 1.5** Calibration modes in analytical sciences. Source: Adapted from Visconti et al. [13].

its concentration is obtained by extrapolation where $Y$-value is equal to 0. This other protocol is illustrated with the example of Section 2.4.3. When the number of spikes is significant, SAM can also be applied when the calibration curve is polynomial, particularly when high endogenous signals affect the linearity of the response due to detector saturation.

When multiple signal-based detectors, such as MS or DAD, can record several physicochemical properties of the target analyte, more than one SAM calibration curve can be simultaneously acquired for the same working sample. This multiple-response monitoring leads to the possibility of dealing with the H-point standard addition method (HPSAM).

This new procedure is effective to control both proportional and additive biases (defined in Chapter 4), such as matrix interferences and/or detector saturation, when all calibration lines are converging at almost the same $X$-intercept. A comprehensive example is presented in Section 10.1. If the calibration lines are not correctly converging, a revised HPSAM was proposed including chemical modifiers [12]. Figure 1.5 is an attempt to propose a schematic overview of the diverse quantification/calibration strategies described in this chapter.

## 1.5 Some New Quantification Techniques

As stated, this chapter does not aim to give an exhaustive description of all possible quantification modes. However, it is valuable to describe some new insight on

a class of analytical techniques involving MS detection hyphenation, while many novel quantification modes were recently developed thanks to the improvement of modern MS instrumentation. More details are available in a recent review [13]. MS-coupled methods have progressively emerged as a one of the key instrumental components for numerous applications in laboratories, thanks to the development of new instruments and the reduction of costs.

The latter has become possible due to the advent of atmospheric pressure ionization interfaces, allowing to produce gas-phase ions that can be further analyzed. Compared to traditional spectroscopic detectors, such as UV absorbance, mass spectrometers offer additional selectivity by determining the mass/charge ratio of ion(s) or transition. An increasing number of articles reporting new MS-coupled methods for quantification are submitted each year [14].

In the field of MS-coupled methods, the greater availability of SILs opens the possibility of novel calibration procedures. They can mainly be employed as ideal surrogate calibrants to directly perform the calibration in the study matrix. Obviously, if they are used for this purpose, the analyst must first investigate the potential presence of interferences with the authentic analyte. When a contribution coming from the SIL is significant and modifies the signal, the application of isotope pattern deconvolution (IPD) was proposed as a corrective approach. In the absence of significant interference, internal calibration represents one of the most promising methodologies for modern absolute quantification.

### 1.5.1 Isotopic Pattern Deconvolution (IPD)

Isotope dilution mass spectrometry (IDMS) is a well-known technique applicable both to organic as inorganic analysis. It is because all isotopes of one element show almost the same chemical properties but mass differences between isotopes that IDMS allows quantifying the analyte by mass spectrometry. There are different IDMS operating procedures offering also various levels of precision. In many routine applications simple and fast operating procedures can be applied. The IPD is one of these high precision procedures based on the natural isotopic pattern alteration of a standard using a minor isotope labeled analog. In contrast to traditional analytical methods that rely on signal intensity, IPD is established by ratioing the signals between the isotopes of the molecule of interest and an analog with an enriched isotopic composition (i.e. SIL).

The IPD is sometimes claimed to be one of the most reliable and highest-quality metrological methods and is commonly used by chemical manufacturers to calculate SIL isotopic enrichment and purity. The isotopic abundance and concentration of the isotope labeled analog can be obtained by reverse isotope dilution mass spectrometry, i.e. a calibration against a high purity solution of the natural analyte prepared from a gravimetric solution of a suitable reference material.

First, the isotopic distributions for unlabeled standard and SIL as well as their combinations are computed using dedicated software: this is the convoluted isotope distribution. Free-access software is available coded with R to achieve the deconvolution. The labeled compound is then added to the reference material, resulting in

isotopic dilution. Then, the comparison between theoretical and experimental isotope overlap allows us to determine the SIL isotopic enrichment, chemical purity, and concentration.

Finally, once the SIL solution has been characterized by isotope dilution mass spectrometry it can be used as a calibrant for IPD quantification [15]. The more detailed procedural aspect is as follows, where variable $A$ is the measured isotopic abundance, subscript *nat* comes for natural, *lab* for labeled and *mix* for mixed.

- Step 1. The natural isotopologue distributions of the analyte $X$ and its isotope labeled analog $X_{SIL}$ are measured. Let us remember that isotopologues only differ in their isotopic composition and have the same chemical formula. Superscripts $M0$, $M1$, etc. used in following formulas to indicate isotopologues.
- Step 2. Authentic analyte and SIL are mixed, and the resulting isotope pattern are determined. The basic concept is to say that the pattern of mixed solution is a linear combination of natural and labeled patterns weighted by the molar fractions $q_{nat}$ and $q_{lab}$, respectively:

Deconvolution model for IPD

$$A_{mix} = q_{nat}A_{nat} + q_{lab}A_{lab} + E \tag{1.5}$$

The vector of random error $E$ is added to account for the errors in the isotopic determinations. It is called a deconvolution model because it is slightly different of the classic calibration model, such as Equation (2.6), where there is only one predictive variable, the calibrant concentration usually noted $X$ as explained in Sector 2.2. In this case there are two predictive variables $A_{nat}$ and $A_{lab}$. Once the isotopic abundances are measured, we have a set of equations:

Isotopic patterns

$$\begin{aligned}
A_{mix}^{\mathrm{M0}} &= q_{nat}A_{nat}^{\mathrm{M0}} + q_{lab}A_{lab}^{\mathrm{M0}} + E^{\mathrm{M0}} \\
A_{mix}^{\mathrm{M1}} &= q_{nat}A_{nat}^{\mathrm{M1}} + q_{lab}A_{lab}^{\mathrm{M1}} + E^{\mathrm{M1}} \\
A_{mix}^{\mathrm{M2}} &= q_{nat}A_{nat}^{\mathrm{M2}} + q_{lab}A_{lab}^{\mathrm{M2}} + E^{\mathrm{M2}} \\
&\cdots \\
A_{mix}^{\mathrm{Mn}} &= q_{nat}A_{nat}^{\mathrm{M}n} + q_{lab}A_{lab}^{\mathrm{M}n} + E^{\mathrm{M}n}
\end{aligned}$$

They can be rewritten in a more condensed matrix form (the term matrix is used with its mathematical meaning) clearly showing this a multiple regression model with two variables and no intercept:

Multiple regression model

$$\mathbf{A}_{mix} = [\mathbf{A}_{nat}\mathbf{A}_{lab}]\mathbf{q}^{-1} + \mathbf{E} \tag{1.6}$$

- Step 3. Apply least-squares multiple linear regression to get the solutions of model 1.6; i.e. the estimates of the molar fractions $q$. With Excel this can be achieved using the `LINEST` built-in function. This function usage is described in Section 2.3.1. In this chapter, `LINEST` is applied to estimate the three coefficients of a quadratic model. Model in equation 1.6 is also a 3-coefficient model, with one coefficient equal to 0.

– Step 4. Knowing the SIL concentration, noted $X_{SIL}$ or equivalently $X_c$, used for spiking the working sample, direct quantification of the analyte $Z$ can be provided without the need for a calibration curve as shown in Equation (1.7).

Estimated sample concentration

$$Z = X_{SIL} \times \frac{q_{nat}}{q_{lab}} \tag{1.7}$$

To correctly achieve this procedure and be able to perform the deconvolution, it is essential to have a crosstalk or isotopic overlap. This is possible when SIL chemical purity and/or isotopic enrichment is less than 100% or when there is only a small mass-unit difference between the isotope labeled standard and its analogous compound. IPD reproducibility was estimated based on an interlaboratory study, including four different World Anti-Doping Agency (WADA) accredited laboratories, and compared to a more traditional EC calibration method using surrogate standards. More details on interlaboratory precision parameters are available in Section 3.1.

The IPD shows the same accuracy and demonstrates improved reproducibility at low concentrations (2 ng/ml) with a relative standard deviation of reproducibility ranging approximately from 10 to 16%, respectively [16]. This result shows that isotope dilution mass spectrometry determination analytical methods are of high metrological quality. To confirm the high metrological reliability of the IPD, MU was estimated the same manner it is presented in Section 6.4 for LEAD example.

Uncertainty budget shows that MU is mainly dependent on the experimental determination of isotopic abundance (78.0%) and SIL concentration measurement (21.3%). Reducing these two sources of uncertainty involves some additional work during method development, but the IPD procedure then benefits from a better performing and faster analysis because calibration is performed within the working sample, and no traditional EC curve is required.

### 1.5.2 Direct Internal Calibration with Labeled Calibrant (IC-SIL)

When possible, the simplest quantification procedure is probably achieved when an in-sample single amount of surrogate calibrant is used to compute the working sample concentration. With this procedure, authentic and surrogate standards are simultaneously measured. The estimated analyte concentration, $Z$, is directly obtained via the peak area ratio of the sample *versus* the surrogate calibrant. Because only one concentration level is introduced in the working sample, a response factor relationship must be first established to confirm the absence of ionization competition between surrogate and authentic analyte, independent of the concentration. Thus, equimolar mixtures of surrogate and authentic analyte in neat, artificial and/or depleted matrices are first analyzed over the investigated calibration range. Additionally, ionization competition at nonequimolar concentrations should be investigated. Thus, several multipoint calibrations using the authentic analyte with surrogate standard at different concentration levels can be analyzed to study the authentic analyte response function alteration. Once the *RF* has been empirically determined, the working sample concentration is calculated as follows:

Working sample concentration

$$Z = \frac{Y}{Y_{SIL}} \times \frac{X_{SIL}}{RF} \tag{1.8}$$

This equation is a reorganization of Equation (1.3), where the intercept is zero, and the slope $a_1$ corresponds to the *RF*. When SIL is spiked at low concentration, such as 12.5 or 25% of the ULOQ, marked competitive ion suppression occurs due to the concomitant presence of the analyte at higher concentrations in ionization source. Conversely, when the SIL concentration is fixed in the highest bound of the response function, the influence of the surrogate signal on a low concentrated analyte can be detrimental and generate a significant bias. A correction procedure was proposed by determining the SIL concentration equivalent, noted $X^*_{SIL}$ obtained with the following formula:

SIL concentration equivalent

$$X^*_{SIL} = X_{SIL} \times P_{SIL} \times E_{SIL} \times \frac{MW_A}{MW_{SIL}}$$

where *MW* is the molecular mass of authentic analyte and SIL surrogate, *P* the chemical purity as percentage, and *E* the isotopic enrichment, expressed as the probability of finding a labeled atom at any single site [17]. New reagents and improved instrumentation give opportunities to develop novel and faster quantification procedures exhibiting high metrological quality parameters. For instance, the one-point calibration method using SIL as calibrant and their isotopes was introduced to extend the lower limit of quantification (LLOQ).

To perform this analysis, a triple quadrupole instrument was used and a particular acquisition method named multiple isotopologue reaction monitoring (MIRM) was developed. By monitoring the SIL isotopic fragmentation abundances, a regression model was constructed by plotting the surrogate standard concentration equivalent on the abscissa and the instrument response (peak areas) of the corresponding MIRM channel on the ordinate. Then, the authentic analyte concentration can be calculated using the regression parameters [18]. This is just an example of regularly active literature.

Overall, internal calibration with SIL as calibrant is conceptually straightforward for absolute quantification with modern MS instrumentation, but requires additional steps during method development, such as the experimental determination of the *RF* and, with the MIRM procedure, isotopic abundance determination. However, once the method is developed, it is markedly faster in routine analysis because a daily repeated calibration curve is no longer required, and comparable results to EC can be obtained. Currently, the IC is raising interest due to the increased number of high-quality SILs commercially available, even if they remain limited to the most classic compounds. To overcome this limitation, isotope standards can be generated in-house by derivatizing authentic analytes with labeled $^{13}C_2$-dansylchloride and $^{13}C_2$-dansylhydrazine.

As an illustration of the selection of quantification procedure, Figure 1.6 presents a flowchart applicable to LC-MS methods. Some parts of this flowchart are transferable to other methods of analysis and/or detection modes. Possible strategies are

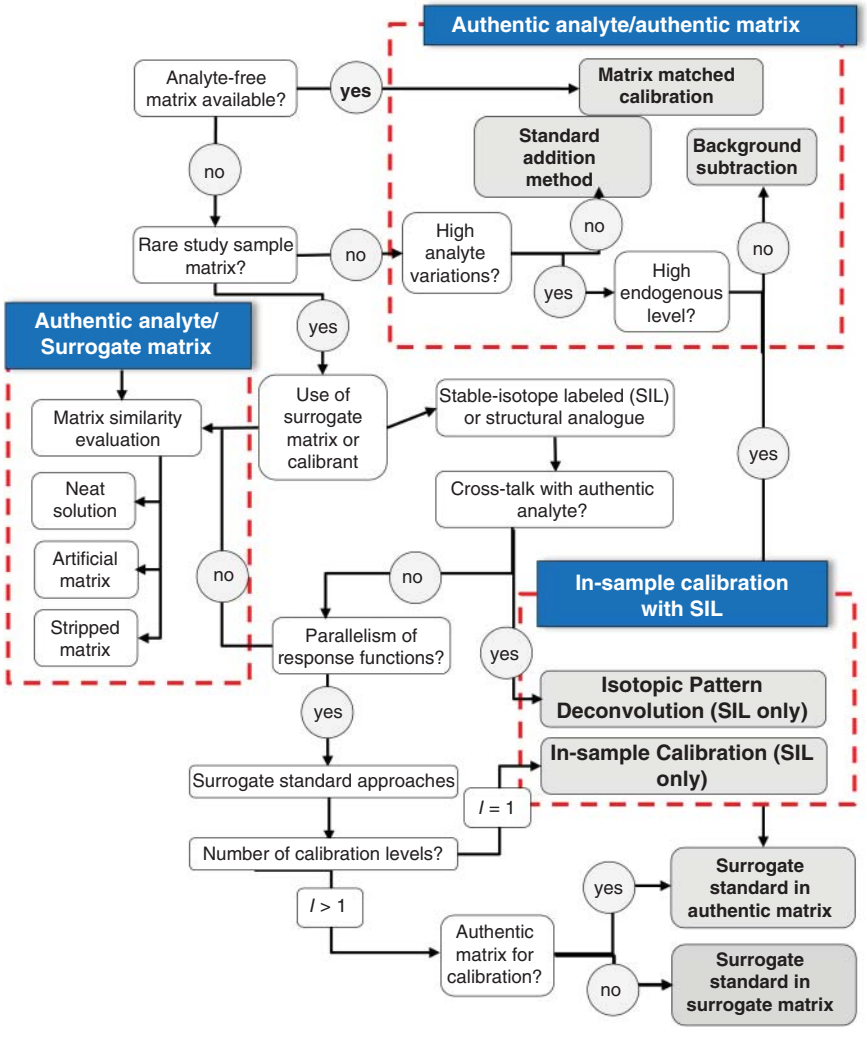



**Figure 1.6** LC-MS on endogenous metabolites: proposed workflow for selecting a calibration operating procedure. Source: Adapted from Visconti et al. [13].

identified by square corner grey boxes, namely, authentic analyte/authentic matrix, authentic analyte/surrogate matrix, and ISC with SIL. For each case, different calibration procedures are appropriate, depending on complementary information about the analyte, the sample matrix availability, or the presence of endogenous analyte in the matrix.

More details are given in [13] and the rest of the chapter. In recent decades, advances in analytical calibration methodologies, instrument technology and enlarged SIL availability have contributed to improving the accuracy and throughput of quantitative analysis. However, the gap in knowledge between published official guidelines and strategies used by the analytical community prevents consensus about exactly how validation should be performed.

The introduction of innovative calibration approaches allowed the analyst to perform the calibration in the authentic working sample matrix, overcoming different bottlenecks such as the lack of blank matrices, the extraction efficiency, and matrix effect between the external calibration curve and unknown samples. Scientific interest is growing around direct internal calibration with SIL due to its analytical process simplicity and quickness to provide quantitative results from a few samples or even a single sample. With these unique advantages, internal calibration strategies have enormous potential to be widely applied for various quantitative applications and may even change the landscape of quantitative analysis, although these methodologies are still not officially endorsed by international guidelines for analytical method validation.

## References


1 BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML (2012). *International Vocabulary of Metrology — Basic and General Concepts and Associated Terms (VIM3)*. JCGM 200. France: Sèvres https://www.bipm.org/ (accessed 23 July 2023).

2 Fanelli, F., Cantù, M., Temchenko, A. et al. (2022). Report from the HarmoSter study: impact of calibration on comparability of LC-MS/MS measurement of circulating cortisol, 17OH-progesterone, and aldosterone. *Clinical Chemistry and Laboratory Medicine* 60 (5): 726–739.

3 Food and Drug Administration (FDA) (2018). *Bioanalytical Method Validation Guidance for Industry*. Washington, DC: Office of Communications, Division of Drug Information Center for Drug Evaluation and Research https://www.fda.gov/files/drugs/published/Bioanalytical-Method-Validation-Guidance-for-Industry.pdf (accessed 31 August 2023).

4 European Medicines Agency (EMA) (2023). ICH Guideline M10 on Bioanalytical Method Validation - Step 5b. EMA/CHMP/ICH/172948/2019. Committee for Human Medicinal Products.

5 Bugner, E. and Feinberg, M. (1992). Determination of mono-and disaccharides in foods by interlaboratory study: quantitation of bias components for liquid chromatography. *Journal of AOAC International* 75 (3): 443–464.

6 Health and Consumer Protection Directorate (DG-SANCO) (2021). Document SANCO No. 11312/2021. *Analytical Quality Control and Method Validation Procedures for Pesticide Residues Analysis in Food and Feed*.

7 International Council on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) (2022). ICH-M10 *Bioanalytical Method Validation And Study Sample Analysis*.

8 Visconti, G., Olesti, E., González-Ruiz, V. et al. (2022). Internal calibration as an emerging approach for endogenous analyte quantification: application to steroids. *Talanta* 240: 123149.

9 Wang, W., Zhuang, X., Liu, W. et al. (2018). Determination of kynurnine and tryptophan, biomarkers of indoleamine 2,3-dioxygenase by LC–MS/MS in plasma and tumor. *Bioanalysis* 10: 1335–1344.

**10** Suhr, A.C., Vogeser, M., and Grimm, S.H. (2016). Isotope inversion experiment evaluating the suitability of calibration in surrogate matrix for quantification via LC–MS/MS—exemplary application for a steroid multi-method. *Journal of Pharmaceutical and Biomedical Analysis* 124: 309–318.

**11** Liigand, P., Liigand, J., Cuyckens, F. et al. (2018). Ionisation efficiencies can be predicted in complicated biological matrices: a proof of concept. *Analytica Chimica Acta* 1032: 68–74.

**12** Wieczorek, M., Rengevicova, S., Świt, P. et al. (2017). New approach to H-point standard addition method for detection and elimination of unspecific interferences in samples with unknown matrix. *Talanta* 170: 165–172.

**13** Visconti, G., Boccard, J., Feinberg, M., and Rudaz, S. (2023). From fundamentals in calibration to modern methodologies: a tutorial for small molecules quantification in liquid chromatography–mass spectrometry bioanalysis. *Analytica Chimica Acta* 1240: 340711.

**14** Seger, C. and Salzmann, L. (2020). After another decade: LC–MS/MS became routine in clinical diagnostics. *Clinical Biochemistry* 82: 2–11.

**15** Castillo, Á., Gracia-Lor, E., Roig-Navarro, A.F. et al. (2013). Isotope pattern deconvolution-tandem mass spectrometry for the determination and confirmation of diclofenac in wastewaters. *Analytica Chimica Acta* 765: 77–85.

**16** Pitarch-Motellón, J., Roig-Navarro, A.F., Sancho, J.V. et al. (2021). Isotope pattern deconvolution as a successful alternative to calibration curve for application in wastewater-based epidemiology. *Analytical and Bioanalytical Chemistry* 413: 3433–3442.

**17** Khamis, M.M., Adamko, D.J., and El-Aneed, A. (2017). Development of a validated LC-MS/MS method for the quantification of 19 endogenous asthma/COPD potential urinary biomarkers. *Analytica Chimica Acta* 989: 45–58.

**18** Gu, H., Zhao, Y., DeMichele, M. et al. (2019). In-sample calibration curve using multiple isotopologue reaction monitoring of a stable isotopically labeled analyte for instant LC-MS/MS bioanalysis and quantitative roteomics. *Analytical Chemistry* 91: 2536–2543.

# 2

# Calibration

## 2.1 Direct and Inverse Calibration

Analytical methods are based on well-known physicochemical or biochemical phenomena which can be described by various equations. To illustrate the practical problems when calibrating a method of analysis, we can refer to techniques widely used in laboratories. In conventional ultraviolet–visible (UV–Vis) spectrophotometry, the ratio of the emitted flux $\Phi_0$ to the transmitted flux $\Phi$ after passing through the solution being measured, or transmittance, is used for instrumental response. Depending on the type of spectroscopic method, the logarithm of the transmittance is called the absorbance, transmittance, or optical density (*OD*). According to Beer's law, this ratio depends on the concentration $X$ of the solution, the length $L$ of the measuring cell, and a molar extinction coefficient $\varepsilon_T$ which depends on the temperature $T$ and, of course, the nature of the analyte giving the well-known Beer's law theoretical model:

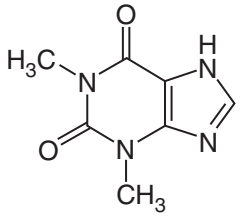
$$\log\left(\frac{\Phi_0}{\Phi}\right) = \varepsilon_T \times X \times L$$

In all cases, the logarithmic transformation of the flux ratio aims to linearize the response function. As explained in Section 1.1.3, this pretreatment is universal and invisible to the analyst. Obviously, it would be sufficient to know the value of $\varepsilon_T$ to predict the responses depending on the values of $X$. Unfortunately, this coefficient varies greatly when the emitted radiation is not purely monochromatic or when the temperature is not constant. In practice, Beer's law cannot be directly used for calibration, since the measuring system is not isolated and is subject to random disturbances due to environmental variations, such as temperature, atmospheric pressure, instability of the electrical voltage, sound vibrations, aging of the equipment components, etc. All these disturbances together form the background noise, which can be estimated, measured, and subtracted.

This simple example shows how calibration must rely on a more complex model than initially imagined from the fundamental equations of physics or chemistry. To illustrate the statistical issues of calibration, we are taking a method of analysis more complex than spectrophotometry, involving mass spectrometry. This chapter is limited to external calibration as it represents the most commonly used calibration

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

mode. Concerning internal calibration or in-sample calibration, several modeling approaches were already presented in Chapter 1.



Theophylline

The same example will be used in different chapters as a roadmap to illustrate the proposed strategy for method validation and measurement uncertainty (MU) estimation. This is a method of analysis of theophylline in human plasma. More details about the complete THEOPHYLLINE method and dataset are given in Table 5.1 and in Section 5.2.2, introducing the method accuracy profile (MAP) procedure.

Theophylline is a molecule used to manage the symptoms of asthma and other lung ill-conditions caused by reversible airflow obstruction. It works by relaxing the smooth muscles in the bronchial airways and pulmonary blood vessels. The aim of the study used for this example was to develop and validate a method for the quantitative determination of different xanthines, including theophylline, caffeine, theobromine, and paraxanthine, in human plasma. This was achieved within the framework of a sports medicine project interested in the effect of these molecules on athletes' performance. These analytes are determined by ultra-high performance liquid chromatography (UHPLC), coupled with a tandem mass spectrometer detection UHPLC-MS/MS [1].

Five calibrators were prepared and contained the analytical grade theophylline (99% purity) at 0.02, 0.1, 0.5, 2.5, and 10.0 µg/l. For each calibrator, two replicates were prepared. As explained in Section 5.2.2, this calibration experimental design was repeated over six different days giving six series of similar calibration datasets. Table 2.1 presents only one of these series or days, called series 1. Other series are comparable and give the same conclusions. The graphical illustration of these data is straightforward when using a worksheet.

The observed points seem to be correctly fitted to a straight line. With this software, it is easy to quickly add on the same graphics several trendlines, such as straight-line, second-order, or quadratic polynomial. When looking at the $r^2$ values reported in

**Table 2.1** THEOPHYLLINE – raw calibration of series 1 dataset, measurement values are expressed in arbitrary units (AU).

| Concentration (µg/l) | Replicate 1 | Replicate 2 |
|---|---|---|
| 0.02 | 0.293 | 0.443 |
| 0.10 | 1.874 | 1.810 |
| 0.50 | 8.904 | 8.306 |
| 2.50 | 23.411 | 37.832 |
| 10.00 | 124.84 | 129.605 |

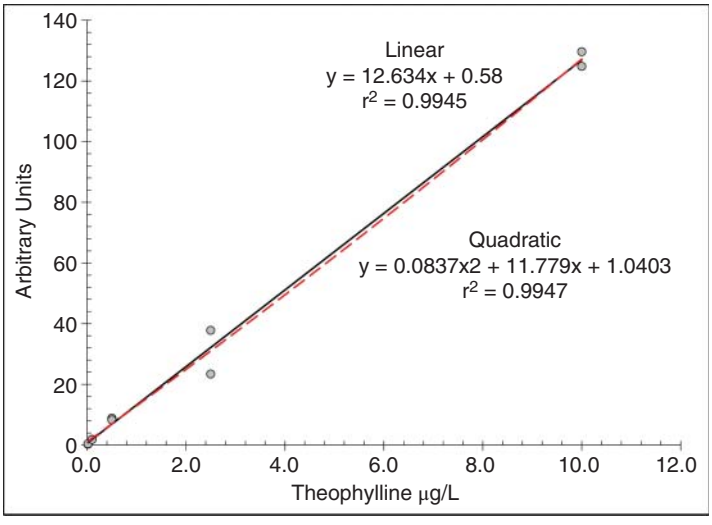



**Figure 2.1** THEOPHYLLINE – illustration of the calibration data of series 1. Solid line: linear model. Dashed line: quadratic model.

Figure 2.1, both models seem equivalent and well-matched. But this simple visual evaluation is misleading, as shown further in Section 8.1. The general statistical model that is used to establish the calibration curve can be described by Eq. (2.1):

Generic direct calibration curve

$$Y = f(X_c) \tag{2.1}$$

- $X_c$ the known concentration of the calibration compound (authentic or surrogate, as explained in Section 1.1.2).
- $Y$ the measured instrumental response.
- $f$ the deterministic model chosen for the response function.

As a reminder, the aim of calibration is not to prove that some relationship between the concentration and the instrumental response exists. All quantitative analysis methods are founded on well-known physicochemical or biological mechanisms, and it is well established that such a relationship exists. In that context, calibration is a two-step procedure:

- Step 1: *Direct calibration*. It consists of collecting instrumental responses of calibrators and calculating the calibration curve coefficients that most accurately relate the (known) calibrant concentration to the instrumental response.
- Step 2: *Inverse calibration*. It uses the inverse function of the calibration model to predict the concentration of unknown samples based on the instrumental response they provide. The values obtained by this operation are called *inverse-predicted* concentrations. Sometimes they can also be called back-calculated, but this may be confusing as the back-calculation procedure is employed in various fields of technology, such as real-time process monitoring, to restore previous situations from historical data. The mathematical interpretation

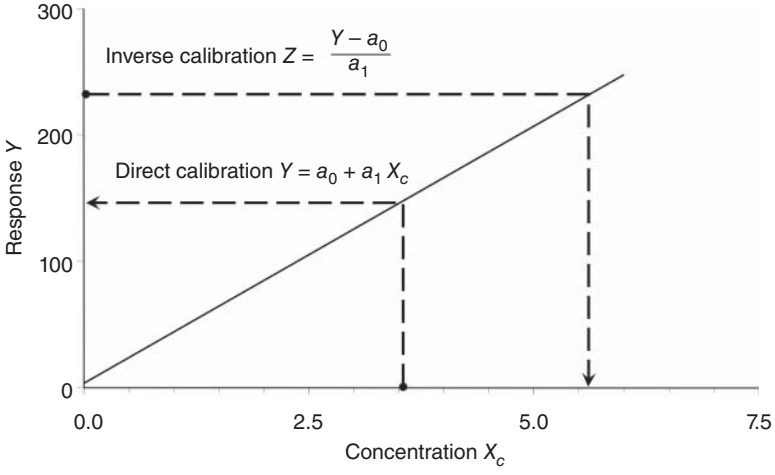



**Figure 2.2** Direct calibration and inverse calibration.

of this step is summarized by Eq. (2.2), which gives its name to the inverse calibration:

Inverse calibration function

$$Z = f^{-1}(Y) \tag{2.2}$$

The two-step quantification procedure is illustrated in Figure 2.2 in the case of a linear calibration model. With theophylline series 1 calibration data, two models, graphically close, can be adjusted, namely linear and quadratic. The question of selecting the best model is raised.

A viable way to answer is to collect new data from samples with known contents, estimate their inverse predicted concentrations and verify which model gives the best inverse predictions. It remains to define what means *the best* calibration model. The role of inverse calibration is quite specific to analytical sciences, and several authors proposed to organize calibration in one stage to avoid this inversion of the calibration curve. The calibration would be named *reverse* calibration as discussed in Section 2.5.

## 2.2 Least-squares Regression Method

### 2.2.1 Straight-line Computation

In the case of multipoint (external) calibration, it is necessary to prepare several calibrators and estimate the calibration model. For simplicity, let us denote $X_i$ the known calibrant concentration of the $i$th calibrator. To fully establish the statistical model of the calibration curve, it is necessary to include the background noise noted $E$ in Eq. (2.1). As stated before, it results from various sources of disturbances, such as

instrument instability or calibrator preparation. If $E_i$ is the random variable accounting for the background noise of the instrumental response $Y_i$ measured on calibrator $X_i$, the observed model is given by Eq. (2.3).

Number of calibration measurements

$$1 \leq i \leq I$$

Measured instrumental response

$$Y_i = f(X_i) + E_i \tag{2.3}$$

Predicted instrumental response

$$\hat{Y}_i = f(X_i) \tag{2.4}$$

Residual

$$E_i = Y_i - \hat{Y}_i \tag{2.5}$$

A part of the model is deterministic and corresponds to $f(X)$ the other is random. The determinism is related to the physicochemical or biological phenomena used to explain that $Y$ is modified when $X$ varies. To make clear this dual structure of the model, statisticians introduced the notation $\hat{Y}_i$ (hat) corresponding to a predicted value of $Y$ once $f$ is known. The difference between measured and predicted instrumental responses corresponds to the random part, i.e. unpredictable, of the model and is called the residual. It is defined by Eq. (2.5).

In Figure 2.3 the linear model is used to illustrate these concepts and explain how the least-squares algorithm works for estimating the coefficients of the $f$ function but it can be transposed to any other model. The observed linear model is given by Eq. (2.6). In this case, the intercept $a_0$ is interpreted as the blank, and the slope $a_1$ as the sensitivity of the method. Knowledge of these coefficients is essential to perform the inverse calibration stage.

Observed linear calibration curve

$$Y_i = a_0 + a_1 X_i + E_i \tag{2.6}$$

Theoretical linear calibration curve

$$\hat{Y}_i = a_0 + a_1 X_i \tag{2.7}$$

The ordinary least-squares (OLS) method is the most widespread algorithm for estimating $a_0$ and $a_1$. A major assumption so that OLS can satisfactorily apply is that the two variables $Y$ and $X$ must play different roles. $Y$ is explained or dependent, while $X$ is said to be explanatory or independent. This means there is a causal relationship between the two variables. Because the instrumental response $Y$ is fully explained by the variations of the concentration $X$, calibration is a typical example for least-squares application. This remark impacts the nature of the residual random variable $E_i$. In practice, it is realistic to suppose it is formed of two components: $E_Y$ linked to the instrumental background and $E_X$ linked to the errors of preparation of the calibration solutions.

$$E_i = E_X + E_Y$$

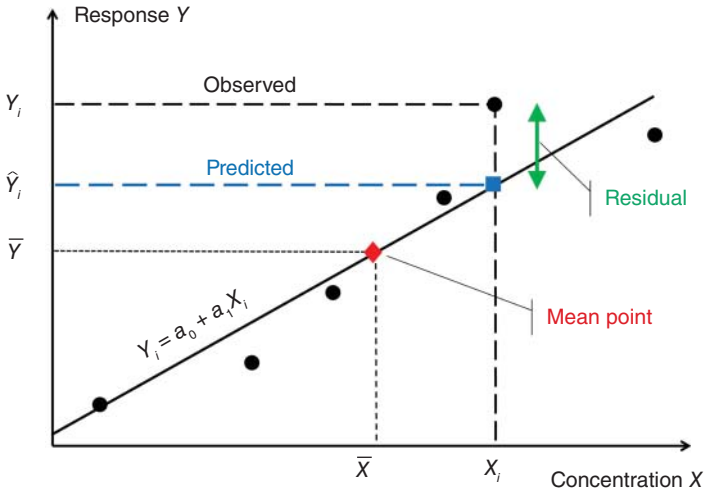



**Figure 2.3** Principles of ordinary least-squares (OLS) method.

But, to apply the OLS method, it must be assumed that $E_X$ is negligible compared to $E_Y$ and explicitly set to zero ($E_X = 0$). In other words, it is assumed that calibrators are prepared with no error. As explained in Section 6.6.2, it is manageable to estimate the MU for each calibrator and verify if it is null or negligible. Given the way calibrators are usually prepared, this assumption seems reasonable as a first approximation. The major downside is that it is always possible to apply OLS whenever the basic assumption is wrong, but the interpretation of the obtained coefficients may be erroneous.

The principle of the OLS method is to estimate the coefficients $a_0$ and $a_1$ by minimizing the sum of the squared differences between observed and predicted values, called residuals, or also deviates, and expressed by Eq. (2.5). Such a difference can be geometrically interpreted as a distance symbolized by a double green arrow in Figure 2.3. The first idea could be to calculate the average of residuals and try to make it as small as possible. Intuitively, it is easy to understand that any line shall pass by the mean point reported as a red diamond on the graphics.

Unfortunately, if the selected line regularly passes through all points, the simple average distance shall always be zero because the sum of negative residuals shall be exactly balanced by that of positive residuals. To avoid this inconvenience, the residuals are generally squared. This numerical solution gave its name to the method: least-squares regression. Other solutions exist, such as summing the absolute differences, but are not as common as the least-squares algorithm. It can be summarized by the following equations:

Residual

$$E_i = Y_i - \hat{Y}_i \tag{2.5}$$

Condition on the sum of squared residuals

$$S = \sum_{i=1}^{I} E_i^2 \rightarrow \text{Minimum} \tag{2.8}$$

In the case of a straight-line, Eqs. (2.6) and (2.7) are combined to obtain the residual, and Eq. (2.8) is modified as follows:

The sum of squared residuals (for the straight line)

$$S = \sum_{i=1}^{I} (Y_i - a_0 - a_1 X_i)^2 \rightarrow \text{Minimum} \tag{2.9}$$

Equation (2.9) is a function with two unknowns $a_0$ and $a_1$ since all values $X_i$ and $Y_i$ are known. A function passes through minimum at the point where the first derivative is zero. The first two derivatives, one with respect to $a_0$ and the other with respect to $a_1$ are calculated and set to 0. A set of simultaneous equations with two unknowns is obtained and solved. After simplification, the calculation comes down to the following set of formulas directly applicable while all elements are known. Notation *SS* for sum of squares and *SP* for sum of crossed products are used for simplicity.

Mean point

$$\overline{X} = \frac{\sum_{i=1}^{I} X_i}{I} \quad \overline{Y} = \frac{\sum_{i=1}^{I} Y_i}{I} \tag{2.10}$$

Sum of crossed products of deviates

$$SP(X,Y) = \sum_{i=1}^{I} (X_i - \overline{X})(Y_i - \overline{Y}) \tag{2.11}$$

Sum of squared deviates for $X$

$$SS(X) = \sum_{i=1}^{I} (X_i - \overline{X})^2 \tag{2.12}$$

Slope or sensitivity

$$a_1 = \frac{SP(X,Y)}{SS(X)} \tag{2.13}$$

Intercept or blank

$$a_0 = \overline{Y} - a_1 \overline{X} \tag{2.14}$$

Variance of the residuals or residual variance

$$\begin{cases} s_E^2 = \dfrac{\sum_{i=1}^{I} (Y_i - \hat{Y}_i)^2}{I-2} \\ s_E^2 = \dfrac{SS(Y) - a_1 SP(X,Y)}{I-2} \end{cases} \tag{2.15}$$

Residual standard deviation

$$s_E = \sqrt{\frac{SS(Y) - a_1 SP(X,Y)}{I-2}} \tag{2.16}$$

### 2.2.2 Assumptions and Complements

These results are established under the following set of assumptions:

- Variables $E_i$ are randomly distributed according to Normal laws $\mathcal{N}(0, \sigma^2)$ with zero means and theoretical variances $\sigma^2$. All laws associated with $E_i$ are identical throughout the calibration domain. This means all response variances for all calibrators (or standard solutions) are constant and equal to $\sigma^2$. The best estimate of $\sigma^2$ is equal to $s_E^2$ (Eq. 2.15). If this assumption is not acceptable, a modified method called weighed least-squares (WLS) must be used, as explained in Section 2.3.2.
- Variables $E_i$ are independent, which means that they are not correlated. To comply with this assumption, it is recommended to separately prepare the calibrators. The common incorrect practice of making successive dilutions can exacerbate this dependence and must be discouraged. Despite this well-known tip, many commercial kit manuals still advise making successive dilutions.
- Outliers must be eliminated beforehand. The OLS method, far from pointing out model errors or nonconforming values, will force the model to pass through all points. Moreover, the use of a sum of squares as a minimization criterion gives much greater weight, or leverage, to any point which is far from the others and can thus very easily introduce a bias; it is not a robust method.

When these assumptions are considered acceptable (although rarely verified), it is then possible to calculate various additional parameters described by Eqs. (2.17)–(2.22). The residual standard deviation $s_E$ allows to obtain other standard deviations, such as the blank, the sensitivity, or a predicted value $\hat{Y}$ and the associated confidence intervals. Ultimately, $s_E$ gives a global indication of the fitting closeness between the predicted line and the experimental points.

Standard deviation of sensitivity

$$s_{a_1} = \frac{s_E}{\sqrt{SS(X)}} \tag{2.17}$$

Standard deviation of analytical blank

$$s_{a_0} = s_E \sqrt{\left(\frac{1}{I} + \frac{\overline{X}^2}{SS(X)}\right)} \tag{2.18}$$

Confidence interval of sensitivity

$$\left[a_1 \pm t_{1-\frac{\alpha}{2};I-2} \times s_{a_1}\right] \tag{2.19}$$

Confidence interval of blank

$$\left[a_0 \pm t_{1-\frac{\alpha}{2};I-2} \times s_{a_0}\right] \tag{2.20}$$

Quantile of Student's t for a confidence level of $1-\alpha$, $I-2$ degrees of freedom

$$t_{1-\frac{\alpha}{2};I-2}$$

Predicted response for a given $X$ value

$$\hat{Y} = a_0 + a_1 X \tag{2.21}$$

Standard deviation of a predicted response

$$s_{\hat{Y}} = s_E \sqrt{\left( \frac{I+1}{I} + \frac{(X_k - \overline{X})^2}{SS(X)} \right)} \tag{2.22}$$

The confidence intervals as given by Eqs. (2.19) and (2.20) are applied in Section 2.4.3 to present a simplified method for verifying the parallelism of two straight lines. Confidence intervals and tests of hypothesis are procedures for making so-called statistical inferences, i.e. the domain of statistics devoted to making global decisions, or inferring a conclusion, from limited information obtained on a sample. It is considered that hypothesis tests are seldom needed if confidence intervals are available because they are similarly effective and can be graphically interpreted. When two confidence intervals are overlapping, this means that the statistical parameters are not significantly different, or when a confidence interval contains a given value, such as 0, the related parameter is not significantly different from this value.

Up to now, least-squares algorithm is described for the straight-line case, i.e. the model containing one first-order explicative variable denoted $X$. Multiple regression models can also be handled by OLS and similar formulas obtained. To demonstrate how the OLS algorithm works in the case of more complicated multivariate models, it requires using mathematical matrix notation and calculus. An example is presented in Section 1.5.1, even if this is out of the scope of this book. Full description is available in many statistical textbooks, e.g. the excellent reference book of Draper and Smith about regression [2].

In the next chapter, an example of the quadratic model is presented, and it shows how the coefficients can easily be computed with Excel. Once the coefficients are estimated, the inverse calibration function can be used to obtain the inverse-predicted concentration of unknown samples. Let us denote $Y_k$ the instrumental response recorded with an unknown sample $k$ and $Z_k$ its inverse-predicted concentration given by Eq. (2.24) in the case of a straight-line:

Inverse – predicted concentration (general model)

$$Z = f^{-1}(Y) \tag{2.23}$$

Inverse – predicted concentration (linear model)

$$Z_k = \frac{Y_k - a_0}{a_1} \tag{2.24}$$

As explained in Section 1.1.1, to simplify notation but emphasize the difference between $X_c$ the concentration in the calibrator is controlled by the experimenter and $\hat{X}$ the concentration of the authentic analyte predicted for an unknown sample, the inverse-predicted concentration is denoted $Z$ in the following chapters. However, a more straightforward model for $Z$ could be $Z = g(Y)$. This model defines what can be called *reverse* calibration. It would mean that calibration is no longer a two-step procedure. The inverse calibration function becomes useless to predict the concentration of unknown samples, whereas reverse prediction model could directly accomplish this.

The pros and cons of the reverse model are explained in Section 2.5. Depending on the quantification procedure, the principal benefit of this notation is to emphasize the idea that it is not always obvious that $g = f^{-1}$. It must also be noted that the standard deviation of a predicted response $s_{Y_k}$ given by Eq. (2.22) must not be confounded with the standard deviation $s_Z$ of the inverse-predicted concentration. The basic reason is that $s_{Y_k}$ is related to $Y$ and expressed in the same units as the instrumental response, while $s_Z$ is related to the concentration $X$ and in the analyte unit. The possible estimation of this important standard deviation is addressed in Section 2.5.1 as well as the diverse problems raised.

## 2.3 Software Implementation

### 2.3.1 Ordinary Least-squares (OLS) Regression

With any worksheet, the coefficients of the model described by the Eq. (2.6) can be estimated by using the built-in function `LINEST`. An example of application to the data of Table 2.1 is presented in the worksheet named Resource A.

Some explanation will help in understanding this worksheet. Beforehand, calibration data of Table 2.1 must be unfolded into two columns and stored in cells `A5:B14`. In cell `A16`, the built-in function `=LINEST (B5:B14; A5:A14;TRUE;TRUE)` is typed. Results returned by `LINEST` are illustrated below in the case of a simple straight-line regression with two coefficients. To make the worksheet easily recycled by any analyst, formulas are visible in the column on the right of the cell, giving the result. For instance, the formula applied in cell `A22` is visible in cell `C22`. The interpretation of the results appearing in range `A16:B20` is more complex, as explained below.

Following Excel nomenclature, `LINEST` is a matrix (or tabular) function, i.e. it requires one or more matrices as arguments and returns several results as a matrix (in the mathematical meaning of the word). To exactly understand the way arguments are entered and results returned, it is necessary to look at the user's manual.

| | a1 | a0 | |
|---|---|---|---|
| Coefficients | 12.634 | 0.580 | |
| $s_{a_1}$ | 0.332 | 1.530 | $s_{a_0}$ |
| $r^2$ | 0.995 | 3.981 | $s_E$ |
| $F$-value | 1452.238 | 8 | Degrees of freedom |
| Regression $SS$ | 23011.141 | 126.762 | Residual $SS$ |

According to the Excel version, the procedure to apply matrix functions could differ. In recent releases, typing the function in one cell allows us to obtain the results automatically distributed in a set of cells. If there is no place for this operation, an error is emitted. In older versions, it was necessary to first select the destination range and simultaneously type three keystrokes Shift, Ctrl, and Enter. Whatever the

method, in cells `A15:B19`, results are arranged as follows (*SS* stands for Sum of Squares):

**Resource A Linear and quadratic calibration (Excel).**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Resource A : Linear and quadratic calibration using OLS** | | | | | |
| 2 | Theophylline Series 1 (µg/l) | | | | | |
| 3 | Linear model | | | Quadratic model | | |
| 4 | X | Y (AU) | | X | $X^2$ | Y(AU) |
| 5 | 0.02 | 0.293 | | 0.02 | 0.0004 | 0.293 |
| 6 | 0.02 | 0.443 | | 0.02 | 0.0004 | 0.443 |
| 7 | 0.10 | 1.874 | | 0.10 | 0.0100 | 1.874 |
| 8 | 0.10 | 1.810 | | 0.10 | 0.0100 | 1.810 |
| 9 | 0.50 | 8.904 | | 0.50 | 0.2500 | 8.904 |
| 10 | 0.50 | 8.306 | | 0.50 | 0.2500 | 8.306 |
| 11 | 2.50 | 23.411 | | 2.50 | 6.2500 | 23.411 |
| 12 | 2.50 | 37.832 | | 2.50 | 6.2500 | 37.832 |
| 13 | 10.00 | 124.835 | | 10.00 | 100.0000 | 124.835 |
| 14 | 10.00 | 129.605 | | 10.00 | 100.0000 | 129.605 |
| 15 | | | | | | |
| 16 | 12.634 | 0.580 | =LINEST(B5:B14;A5:A14;TRUE;TRUE) | 0.084 | 11.779 | 1.040 |
| 17 | 0.332 | 1.530 | | 0.188 | 1.949 | 1.915 |
| 18 | 0.995 | 3.9806 | | 0.995 | 4.1963 | #N/A |
| 19 | 1452.238 | 8 | | 653.499 | 7 | #N/A |
| 20 | 23011.141 | 126.762 | | 23014.642 | 123.2614 | #N/A |
| 21 | | | | | | |
| 22 | Coeffcient a0 | 0.580 | | Coeffcient a0 | 1.040 | |
| 23 | Coeffcient a1 | 12.634 | | Coeffcient a1 | 11.779 | |
| 24 | Residual std dev | 3.98061 | | Coeffcient a2 | 0.084 | |
| 25 | sa0 | 1.530 | | Residual std dev | 4.1963 | |
| 26 | sa1 | 0.332 | | sa0 | 1.915 | |
| 27 | $r^2$ | 0.9945 | | sa1 | 1.949 | |
| 28 | | | | sa2 | 0.188 | |
| 29 | | | | $r^2$ | 0.9947 | |

When looking at the returned parameters, the straight-line model seems satisfactory, while the coefficient of determination $r^2$ is extremely high. This parameter, expressed as a percentage, can be interpreted as the percent of the variation of $Y$ explained by the regression model (see Section 2.4.2). The difference with a perfect model corresponds to the amount of unexplained variation. In the example, about 99.5% of $Y$ can be explained by the selected regression model, and 0.5% is unexplained. The opposite would have been surprising because the analytical method was purposedly developed on this physicochemical property.

In the second line of the result matrix, the standard deviations of coefficients are also returned and can be used for further calculations. More details about $r^2$ are available in Section 2.4.2.

On the same worksheet, a second-order polynomial model is adjusted to the same data. Equation (2.25) specifies the quadratic model used in this example and the corresponding inverse calibration function. The three coefficients are also estimated by using built-in function `LINEST` with the arguments `=LINEST(F5:F14;D5:E14;TRUE;TRUE)`.

Quadratic calibration model

$$Y_i = a_0 + a_1 X_i + a_2 X_i^2 + E_i \tag{2.25}$$

Inverse quadratic calibration model

$$Z_k = \frac{-a_1 + \sqrt{a_1^2 - 4a_2(a_0 - Y_k)}}{2a_2} \tag{2.26}$$

It is well-known that the second-order equation has two workable solutions, similarly, the inverse calibration quadratic model may come down to two values. The solution presented in Eq. (2.26) assumes that the calibration curve is concave. This is the usual situation when $Y$ is regularly increasing when $X$ is increasing. Hence, to apply `LINEST` to this model with three parameters, a new column containing the squared values of $X$, called "$X^2$" must first be added. The built-in function `LINEST` is very flexible and can also be used for many applications. For instance, it can cope with multiple linear regression models such as employed for the isotopic pattern deconvolution method described in Section 1.5.

It appears in the range `E5:E14`. In cell `D16`, the formula which generates the results is typed. It is not displayed on the worksheet for lack of space, but it is easy to see that both columns `D` and `E` containing $X$ and $X^2$, respectively, are input as the second argument. The coefficient of determination $r^2$ of the quadratic model equals 0.995 and is identical to the linear model. Therefore, this criterion is not sufficient to select the best calibration model, and another approach is needed, as explained in many statistical handbooks [2].

To attempt to define a procedure suitable for the selection of the best model that can be estimated from the calibration data of Table 2.1 collected on one single day, replicate measurements were done on six samples spiked with known amounts of theophylline. Instrumental responses expressed in arbitrary units (AU) are reported in Table 2.2. No measurement is out of the calibration range as the lowest concentration here is 0.05 µg/l, and the lowest calibrator was 0.02 µg/l.

**Table 2.2** THEOPHYLLINE series 1 – validation data.

| | | Inverse-predicted concentrations (µg/l) | | | |
|---|---|---|---|---|---|
| **Spike (µg/l)** | **Response (AU)** | **Linear** | **Bias (%)** | **Quadratic** | **Bias (%)** |
| 0.05 | 1.307 | 0.058 | 15.1 | 0.023 | −55 |
| 0.05 | 1.259 | 0.054 | 7.5 | 0.019 | −63 |
| 0.10 | 1.909 | 0.105 | 5.2 | 0.074 | −26 |
| 0.10 | 1.883 | 0.103 | 3.1 | 0.072 | −28 |
| 0.50 | 8.638 | 0.638 | 27.6 | 0.642 | 28 |
| 0.50 | 8.786 | 0.650 | 29.9 | 0.655 | 31 |
| 1.00 | 18.154 | 1.391 | 39.1 | 1.438 | 44 |
| 1.00 | 17.672 | 1.353 | 35.3 | 1.398 | 40 |
| 2.50 | 39.004 | 3.041 | 21.7 | 3.152 | 26 |
| 2.50 | 37.949 | 2.958 | 18.3 | 3.067 | 23 |
| 10.00 | 123.565 | 9.735 | −2.7 | 9.729 | −3 |
| 10.00 | 126.487 | 9.966 | −0.3 | 9.947 | −1 |

Instrumental responses and inverse-predicted concentrations for six spiked samples.

This comment is made because it is always dangerous to extrapolate. With these new responses, inverse-predicted concentrations were calculated with both inverse models using the estimated model coefficients reported in Resource A and put together in the same table. It is additionally possible to compute the individual bias for each result as described in Section 4.1.1. Obtained bias values are hugely varying and disappointing, while both models seem remarkably similar when looking at Figure 2.1. These results are very unsatisfactory.

The reason for this behavior is that the calibration data do not respect one of the mandatory assumptions for applying the OLS regression: the variance of response $Y$ must be constant over all the calibration range, and it is not the case here. Another algorithm must be applied, known as Weighted Least-Squares or WLS.

### 2.3.2 Weighted Least-squares (WLS) Regression

As explained, the implicit assumption for applying OLS method is that all random variables $E_i$ have the same theoretical variances, denoted $\sigma^2$. As explained before, the best estimate of this variance is the residual variance denoted $s_E^2$ and given by Eq. (2.15). This property, called homoscedasticity, means that the variances of the $Y$-values are homogeneous over all the calibration range. When the calibration range is large, it is not unusual that repeated instrumental responses be more spread out at high concentrations than at low concentrations.

Calibrator measurement dispersion often increases when the concentration increases and response variances are no longer identical throughout the calibration domain. When dealing with MU the relationship between concentration and response dispersion is a key issue. For instance, most LC-coupled methods (e.g., inductively coupled plasma-mass spectrometer [ICP-MS], liquid chromatography-mass spectrometer [LC-MS], etc.) cover a large calibration range, and $Y$-value variances significantly increase with analyte concentration. Because low concentrations have smaller variances, precision at the lower end of the range may therefore be compromised and impair the limits of quantification (LOQ).

The practical answer of statisticians was to develop the theory of WLS regression. The basic idea is to compensate for the variation in dispersion of the response $Y$, by weighting the residuals in the opposite manner to the increase in variability. The most logical weighting is to use the inverse of local variances of $Y$ for each calibrator:

$$W_i = \frac{1}{s_{Y_i}^2}$$

This means it is necessary to make replicate measurements to have pertinent estimates of the variances. This may become rather cumbersome. In the context of calibration, simpler weighting schemes can be implemented to correct this heteroscedasticity issue and ensure a reliable estimation of the regression parameters. More empirical weighting factors are based either on the concentration of the calibrant or the measured response, providing a simpler means of approximation. In this case, several weights $W_i$ are applicable, such as:

$$W_i = \frac{1}{X^2} \text{ or } \frac{1}{\sqrt{X}} \text{ or } \frac{1}{Y^2} \text{ or } \frac{1}{\sqrt{Y}}, \text{ etc.}$$

It is not always easy to select the best-adapted weights $W_i$ but a practical approach is possible by making several trials. To present some details of the method, we have chosen a classic approach by using the inverse of the squared concentration $1/X^2$. A major criticism can be made of this choice:

– When the calibration range is large, the ratio between the weights applied to the smallest calibrator and the highest is enormous. In the example of Table 2.2, the smallest weight is $1/0.02^2 = 2500$, and the highest $1/10^2 = 0.01$. This 25,000-fold ratio may emphasize the importance of smallest concentration at the expense of the highest, which may become meaningless. It can also be assumed it may be of some consequence on the accuracy of inverse-predicted values.

Analysts must be aware of this possible issue because this weighting is often chosen and implicitly implemented in instrument monitoring software. Reasons are easy to understand. Whatever the calibration experimental design, without replicates, the computation is straightforward. The following example is given to illustrate the efficiency of the WLS despite this questionable weighting choice. Unfortunately, the built-in function `LINEST` does not present the weighted variant of the LS method. It is possible, with some programming effort, to build template sheets that provide the expected results, as illustrated in a publication [3]. The following formulas can be used to develop such a template, but it is limited to the straight-line model.

Classic weights

$$W_i = \frac{1}{X_1^2} \tag{2.27}$$

Weighted average of $X$

$$\overline{X} = \frac{\sum_i W_i X_i}{\sum_i W_i} \tag{2.28}$$

Weighted average of $Y$

$$\overline{Y} = \frac{\sum_i W_i Y_i}{\sum_i W_i} \tag{2.29}$$

Sum of squares for $X$

$$SS(X) = \sum_i W_i X_i^2 - \frac{\left(\sum_i W_i X_i\right)^2}{\sum_i X_i} \tag{2.30}$$

Sum of squares for $Y$

$$SS(Y) = \sum_i W_i Y_i^2 - \frac{\left(\sum_i W_i Y_i\right)^2}{\sum_i Y_i} \tag{2.31}$$

Sum of cross-products

$$SP(X,Y) = \sum_i W_i X_i Y_i - \frac{\sum_i W_i Y \sum_i W_i Y_i}{\sum_i W_i} \tag{2.32}$$

Slope or sensitivity

$$a_1 = \frac{SP(X,Y)}{SS(X)} \tag{2.13}$$

Intercept or blank

$$a_0 = \overline{Y} - a_1\overline{X} \tag{2.14}$$

Another opportunity that we recommend to apply WLS to any model is to program a short Python script, such as Resource B. Likewise, in other Python programs, comments are added in plain text to explain how it works.

**Resource B Calibration using OLS and WLS (Python).**

It is necessary to import the complementary packages, mainly `statsmodels.formula.api`, that contains all linear regression functions.

```
import pandas as pd
import statsmodels.formula.api as sm
```

For this example, the THEOPHYLLINE series 1 data are used. If a more general program must be developed, Python allows many possibilities to read external data in many formats.

```
X = [0.02, 0.02,0.10,0.10,0.50,0.50,2.50,2.50,10.00,
     10.00]
Y = [0.293,0.443,1.874,1.810,8.904,8.306,23.411,37.832,
     124.835,129.605]
```

Complementary variables are added for the quadratic model or the weighting of data.

```
X2 = [i*i for i in X] # squares of X
```

Select the weighting. This underlines the simplicity of this choice, while using another weight may be more cumbersome.

```
W = [1/i**2 for i in X] # weights
```

All variables are organized in a data frame. This step is necessary to apply the regression function of Python.

```
df_data = pd.DataFrame({"x":X, "x2": X2, "y":Y})
```

The calibration function is declared in a straightforward way, as a text using the variable text names: `formula = 'y ~ x'`. Thereafter, it is combined with the data frame into a `model` by applying `sm.ols` function. The fitting is achieved by the attribute `fit()`. Results are stored in a structure that contains all necessary information. The attribute `params` extract a lot of data, such as the coefficients of the model and many others.

```
formula = 'y ~ x'
model = sm.ols(formula, df_data)
```

*(Continued)*

```
results = model.fit()
print("Linear OLS ", results.params)
```

The application of the WLS algorithm is straightforward. The `sm.wls` function is used instead of `sm.ols`. It is necessary to pass the variable `w` containing the weights.

```
model = sm.wls(formula, df_data, weights=W)
results = model.fit()
print("Linear WLS ", results.params)
```

The computation of the quadratic calibration curve is linear except for the formula that becomes `formula = 'y ~ x + x2'`. The subsequent steps are the same.

```
formula = 'y ~ x + x2'
model = sm.ols(formula, df_data) # apply OLS algorithm
results = model.fit()
print("Polynomial OLS ", results.params)
model = sm.wls(formula, df_data, weights=W) # apply WLS
algorithm
results = model.fit()
print("Polynomial WLS ", results.params)
```

Table 2.3 summarizes the statistics returned by these few lines of code. The last column, called for Akaike information criterion (*AIC*) is explained further.

The differences between the coefficients calculated using the OLS or WLS algorithm appear small, except for the quadratic model. As previously stated, the coefficient of determination $r^2$ does not bring much information. If the coefficients of the WLS quadratic model are applied to compute the inverse-predicted concentration of the samples of Table 2.2, the following results are obtained: 0.077, 0.074, 0.114, 0.113, 0.534, 0.543, 1.144, 1.113, 2.560, 2.486, 10.424, 10.829. The average bias remains at +15%. It is not very substantial to decide which model is the best. Another method is needed to select the best model, and a proposal based on MU is made in Section 8.1.

**Table 2.3** THEOPHYLLINE series 1 – comparison of results obtained with OLS and WLS methods on the data of the first series.

| Model | Method | $a_0$ | $a_1$ | $a_2$ | $r^2$ | *AIC* |
|---|---|---|---|---|---|---|
| Linear | OLS | 0.5801 | 12.634 | | 0.995 | 57.776 |
| | WLS | 0.0863 | 14.722 | | 0.942 | 40.036 |
| Quadratic | OLS | 1.0403 | 11.779 | 0.0837 | 0.995 | 59.496 |
| | WLS | 0.0489 | 16.312 | −0.4281 | 0.955 | 39.463 |

## 2.4 Calibration: Special Topics

### 2.4.1 Nonlinear Calibration Curve

For statisticians, the linearity of a model is not defined the same way as in other fields of mathematics. A model is "linear in the parameters" when it is an additive polynomial or can be transformed into such a polynomial, whatever the highest power of any explicative variable. For example, the quadratic model $Y = a_0 + a_1X + a_2X^2$ used before is linear because the second order $X2$ variable can be transformed into a new first-order dummy variable $X_2$, containing the squared values of $X$. The model becomes $Y = a_0 + a_1X + a_2X_2$, and the three coefficients can be favorably estimated using `LINEST`.

The straight-line model is remarkably successful in analytical sciences because the coefficients are easily interpretable. If the sensitivity of a method is "the increase in response relative to the increase in concentration," the coefficient $a_1$ can be interpreted as the sensitivity.

Sensitivity

$$\lim_{\delta X \to 0} \frac{\delta Y}{\delta X} = a_1 \tag{2.33}$$

On the other hand, the coefficient $a_0$ can be assimilated into a blank. Depending on the quantification technique, the *blank* can combine one or several following elements:

- The eventual response of the reagents used for the sample preparation.
- The *endogenous* analyte concentration of the sample when the matrix is present.
- The background instrumental response.

For many laboratory instruments, there is also an electrical blank (or auto zero) which allows the display to be reset to zero. When the matrix or the calibrant are absent and when the instrument is stable or does not present any drift, it is possible to confuse the calibration blank and the electrical blank. However, they do not have the same origin. The electrical zero is purely related to the electronics of the instrument, while the calibration blank may be produced by the reagents.

The situation is more complex if calibration is achieved in the presence of matrix containing an endogenous concentration of the analyte, the coefficient $a_0$ is a mixture of instrumental, reagent, and sample responses as explained in Section 1.4. For some analytical techniques, it is also established that it is impossible to have a linear calibration model. In fact, it happens for numerous detection modes when the calibration extends over a wide concentration range.

For example, ICP emission spectrophotometry allows for very wide calibration ranges (three to four orders of magnitude) and often does not have a straight line for calibration curve. In this case, the interpretation of the coefficients is more delicate since the concept of sensitivity, as defined by the Eq. (2.33), is no longer meaningful: the sensitivity varies with the concentration.

For a given method, it is sometimes possible to modify the measured response and concentration to linearize the calibration model, but this modifies the statistical

properties of $X$ and $Y$ variables. For instance, an old practice was to linearize by using variable $X$ and $Y$ transformations, such as probit or log-probit transformation. This procedure is misleading and has major downsides, whereas it modifies the probability function of variables $X$ and $Y$, which greatly degrades the inverse prediction accuracy and increases the MU. It is better to select a nonlinear model and keep it in its original form. This means it is necessary to apply a nonlinear regression method to estimate the coefficients, even if it seems more complicated.

The chemical mechanisms involved in several methods, such as radioimmunoassay or enzyme-linked immunosorbent assay (ELISA) assay, are relatively complex and not always fully clarified. More elaborate calibration models must be selected to obtain accurate result quantification. For example, Robison–Cox proposed early the 4-parameter logistic model (4PL) [4]. The 5-parameter logistic model also exists. Both models can be used to predict the OD of $Y$ of the antigen-bound fraction as a function of the ligand concentration $X$ for several ligand binding methods. But they can also be inverted to predict the concentration of an unknown sample from its instrumental response.

4-Parameter logistic model (4PL)

$$Y = a_2 + \frac{a_1 - a_2}{1 + \left(\frac{X}{a_3}\right)^{a_4}} + E \tag{2.34}$$

Inverse 4PL

$$Z = a_3 \left(\frac{a_1 - a_2}{Y - a_2}\right)^{\frac{1}{a_4}} \tag{2.35}$$

Although the 4PL model is empirical, the parameters can be interpreted in a way that is satisfactory to an analyst:

- $a_1$ the smallest response of $Y$ when $X$ tends to 0 and corresponds to the lower asymptote of the curve.
- $a_2$ the highest response of $Y$ when $X$ tends to infinity and corresponds to the plateau of the curve.
- $a_3$ the concentration of $X$ when $Y$ is halfway between its maximum and minimum.
- $a_4$ the curvature intensity.

The major difference between this model and the linear models is that it cannot be transformed into an additive polynomial as defined before. It is strictly nonlinear "in the parameters." The theory of least-squares nonlinear regression will not be presented here but is well-documented and illustrated in several books [5]. The nonlinear regression algorithms, both unweighted and weighted, are mostly iterative and quite different from OLS and WLS applicable to the linear case. To demonstrate the feasibility of nonlinear regression is now straightforward and accessible using modern computing facilities.

A practical application to the 4-parameter logistic calibration function of Eq. (2.34) is presented. This is an example applied to a dataset collected with an ELISA method developed to determine an interleukin in blood. Because the number of coefficients to be estimated is substantial (4) and the model is more complex, it is recommended to collect more data than for a simple linear model. Data are gathered in Table 2.4, with the concentration $X$ in pg/ml and the

**Table 2.4** ELISA[a] – calibration for the determination of interleukin 6 using nonlinear regression.

| Concentration $X$ (pg/ml) | $Y_{i1}$ | $Y_{i2}$ | $Y_{i3}$ |
|---|---|---|---|
| 3.91 | 0.326 | 0.348 | 0.331 |
| 7.81 | 0.361 | 0.387 | 0.366 |
| 15.63 | 0.430 | 0.458 | 0.442 |
| 31.25 | 0.571 | 0.593 | 0.582 |
| 62.50 | 0.873 | 0.911 | 0.874 |
| 125.00 | 1.380 | 1.402 | 1.419 |
| 250.00 | 2.167 | 2.174 | 2.143 |
| 500.00 | 2.756 | 2.820 | 2.732 |

Response unit is optical density.
a) Unpublished personal data.

calibrators is large $I = 8$, as well as the number of replicates $J = 3$ per calibrator, i.e. altogether 24 calibration measurements are necessary.

The built-in function `LINEST` is not able to handle such a model. It could be possible to set up a template worksheet using the Excel add-in called Solver, but it requires some expertise. The simplest solution consists in developing a short Python script like Resource C. Returned results are stored in two structures, one called `pop` that contains the values of the coefficients: in this example $a_1 = 0.33166$, $a_2 = 3.51081$, $a_3 = 202.923$, $a_4 = 1.33290$. The other called `cov` contains the covariance matrix. The interpretation of nonlinear regression parameters is quite different from OLS or WLS. It is important to check if the different coefficients are not correlated. A strong correlation between two coefficients would indicate that they are redundant and the model is overfitted. In this example all covariances indicate that the coefficients are independent.

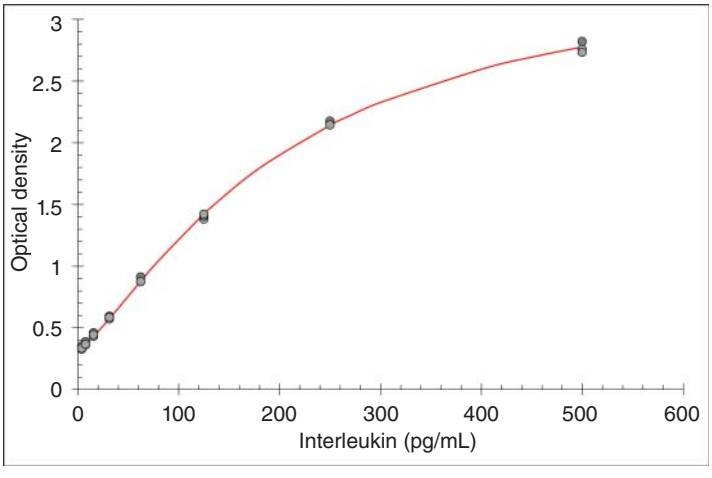



**Figure 2.4** ELISA – determination of interleukin 6. Nonlinear calibration curve and observed measurements.

Figure 2.4 illustrates the ELISA dataset and the continuous line the estimated calibration function. To draw this line, it is necessary to separately compute the predicted response for each concentration using Eq. (2.34). The following short worksheet is a possible application; the column headline "$\hat{Y}$" stands for $\hat{Y}$ and contains some predicted response values.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Nonlinear regression for Interleukin (pg/ml) | | | | | |
| 2 | Coefficients | | | a1 | 0.33166 | |
| 3 | | | | a2 | 3.51081 | |
| 4 | | | | a3 | 202.923 | |
| 5 | | Optical density | | a4 | 1.3329 | |
| 6 | Concentration X (pg/mL) | Yi1 | Yi2 | Yi3 | Y^ | |
| 7 | 3.91 | 0.326 | 0.348 | 0.331 | 0.348 | =$E$3+(($E$2-$E$3)/(1+(A7/$E$4)^$E$5)) |
| 8 | 7.81 | 0.361 | 0.387 | 0.366 | 0.372 | =$E$3+(($E$2-$E$3)/(1+(A8/$E$4)^$E$5)) |
| 9 | 15.63 | 0.430 | 0.458 | 0.442 | 0.433 | =$E$3+(($E$2-$E$3)/(1+(A9/$E$4)^$E$5)) |
| 10 | 31.25 | 0.571 | 0.593 | 0.582 | 0.574 | =$E$3+(($E$2-$E$3)/(1+(A10/$E$4)^$E$5)) |
| 11 | 62.5 | 0.873 | 0.911 | 0.874 | 0.879 | =$E$3+(($E$2-$E$3)/(1+(A11/$E$4)^$E$5)) |

**Resource C Nonlinear calibration (Python).**

For this script, it is necessary to import the `curve_fit` function from the `scipy.optimize` package that performs the nonlinear regression.

```
from scipy.optimize import curve_fit
```

Create a subroutine called `LogisticModel` to declare the 4-parameter of the logistic model. It must be input before any other command line. Beware, the calibration formula is no longer described as a text variable as it was for OLS regression with Python.

```
def LogisticModel(x, a1, a2, a3, a4):
  return(a2 + ((a1 - a2)/(1 + (x / a3)**a4)))
```

Calibration data are stored in local variables but can be input from an external file.

```
X = [3.91,3.91,3.91,7.81,7.81,7.81,15.63,15.63,15.63,
31.25,31.25,31.25,62.5,62.5,62.5,125,125,125,250,250,250,
500,500,500]
Y = [0.326,0.348,0.331,0.361,0.387,0.366,0.43,0.458,
0.442,0.571,0.593,0.582,0.873,0.911,0.874,1.38,1.402,
1.419,2.167,2.174,2.143,2.756,2.82,2.732]
```

The application of the `curve_fit` function is direct. Other arguments can be added, such as initial values of the coefficients when the function may have several optimums. In this example, it is not useful.

```
pop, cov = curve_fit(LogisticModel, X, Y)
```

Print results

```
print(pop) # coefficients
print(cov) # covariance matrix
```

## 2.4.2 Misuses of Regression for Calibration

### 2.4.2.1 Coefficients of Correlation and Determination

Some comments about the use or abuse of the coefficient of determination $r^2$ or its square root, the coefficient of correlation $r$, are imperative. They are often alleged to be efficient criteria to decide whether a calibration model is satisfactory or not. Using proposed notations, calculation formulas are given by Eqs. (2.36) and (2.37):

Coefficient of determination

$$r^2 = \frac{SP^2(X,Y)}{SS(X) \times SS(Y)} \tag{2.36}$$

Coefficient of correlation

$$r = \frac{SP(X,Y)}{\sqrt{SS(X) \times SS(Y)}} \tag{2.37}$$

In this case, $SS(Y)$ represents the sum of the squared deviates of $Y$ to the mean and is calculated as in Eq. (2.12) by replacing the $X$ values by $Y$. The confidence interval of $r$ is not symmetrical and two formulas are necessary to obtain lower $r_L$ and upper limits $r_U$. A prior transformation using the natural logarithm is required.

Initial transformation

$$D = \frac{1}{2} \times \ln\left(\frac{1+r}{1-r}\right)$$

Confidence interval in natural logarithm

$$L = D - \frac{z_{1-\alpha/2}}{\sqrt{I-3}}$$

$$U = D + \frac{z_{1-\alpha/2}}{\sqrt{I-3}}$$

Lower limit

$$r_L = \frac{e^{2L}-1}{e^{2L}+1}$$

Upper limit

$$r_U = \frac{e^{2U}-1}{e^{2U}+1}$$

To obtain the confidence interval, the following lines should be added after the last line 29 of the worksheet Resource A.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 30 | **Coefficent of correlation: Confidence interval** | | | | |
| 31 | Alpha (risk) | 5% | | Alpha (risk) | 5% |
| 32 | Coeff r | 0.997257 | =SQRT(B27) | Coeff r | 0.9973 |
| 33 | Transformation | 3.295231 | =0.5*LN((1+B32)/(1-B32)) | Transformation | 3.309273 |
| 34 | Coverage factor | 1.960 | =NORM.S.INV(1-B31/2) | Coverage factor | 1.960 |
| 35 | Number of points | 10 | =COUNT(A5:A14) | Number of points | 10 |
| 36 | L | 2.5544 | =B33-B34/SQRT(B35-3) | L | 2.5685 |
| 37 | U | 4.0360 | =B33+B34/SQRT(B35-3) | U | 4.0501 |
| 38 | Lower limit | 0.9880 | =(EXP(2*B36)-1)/(EXP(2*B36)+1) | Lower limit | 0.9883 |
| 39 | Upper limit | 0.9994 | =(EXP(2*B37)-1)/(EXP(2*B37)+1) | Upper limit | 0.9994 |

The coefficient of determination of the linear model is $r^2 = 0.9945$ and appears in cell `B27` of the Resource A worksheet, delivered by `LINEST`, and $r^2 = 0.9947$ for the quadratic model in cell `E29`. To obtain the coefficients of correlation, the square roots of the coefficients of determination are extracted in cells `B32` and `E32`, respectively. In this case, they are positive because $X$ and $Y$ variables are always positively correlated. When the correlation is negative, it is compulsory to force to the negative square root. Finally, it appears that the 95% confidence intervals of the coefficient of correlation $r$ of both models are remarkably close, almost overlapping, and equal to [0.9880; 0.9994] and [0.9883; 0.9994], respectively. As already stated, it is impossible to decide which model is the best. For this, some fundamental reasons are helpful.

Historically, $r$ was introduced to measure the correlation, i.e. the proper relationship between two variables. By construction, $r$ value is always between $-1$ and $+1$. When it is positive, it means that the two variables tend to increase simultaneously and are claimed "positively correlated," and in the opposite way, they are "negatively correlated"; when $r$ is close to 0, this means the two variables are "uncorrelated."

A test to check whether $r$ is statistically different from 0 can be assessed or its confidence interval can be used. If the null hypothesis, i.e. $r \neq 0$, is accepted, it is concluded that the two variables are correlated as in calibration. The coefficient of determination $r^2$ is always positive, and varies between 0 and 1. Expressed as a percentage, it was demonstrated it can be construed as the percentage of the variation of $Y$ explained by the regression model. The complement $1 - r^2$ can also be used to account for what is not explained by the model. An adjusted coefficient of determination also exists weighed by the number of points used for computing.

In the case of calibration, the correlation between the response and the concentration cannot be questioned since it is the mechanism that originates the analytical method principle. Therefore, both coefficients $r$ and $r^2$ must be extremely high and (fortunately) always statistically different of zero. A set of arguments show that the use of $r$ or $r^2$ is not relevant to assess the adequacy of the calibration model as underlined some years ago in analytical literature [6, 7].

Cited papers and others also explain why these coefficients are not suitable for selecting the best calibration model and must be avoided. They can be used for quality control (QC) purposes and to check that no calibration measurement is susceptible to be outlier. To summarize, there is no test to substantiate the hypothesis $r = 1$ and, by construction, the confidence interval of $r$ can never include the value 1.

Therefore, it cannot be concluded that slight variations around the $r$ value are representative of the adequacy of the calibration model. The same conclusions are applicable to $r^2$. For instance, the results compiled in Table 2.3 indicate that the coefficient of determination. $r^2$ cannot be used to decide whether the best calibration model is linear or quadratic.

The *AIC* "Akaike Information Criterion (AIC)" is less well known, but it is a statistical parameter specially developed for model selection, i.e. to compare

different possible models and decide which one best fits the data. The best-fit model, according to *AIC* is the one that explains the greatest amount of variation using the fewest possible independent variables:

Akaike information criterion

$$AIC = 2K - 2\ln(L)$$

where:

- $K$ the number of independent variables used to build the model.
- $\ln(L)$ the estimate of the logarithm of the maximum likelihood of the model (how well the model reproduces the data).

The default value of $K$ is always 2, so if the model uses one independent variable $K$ will be 3, if it uses two independent variables, $K$ will be 4, and so on. To compare models, it requires calculating the *AIC* of each model. If a model is more than 2 *AIC*-units lower than another, it is considered significantly more appropriate. Referring to Table 2.3, the best calibration model would be the second-order polynomial estimated by WLS regression. In the Section 8.1 devoted to the accuracy profile, a more empirical strategy to select the best calibration model is developed.

#### 2.4.2.2 Definitions of Linearity

There is also some confusion about the linearity often associated with method validation. In fact, linearity can receive two different definitions:

- The ability to obtain an instrumental response $Y$ strictly related to the concentration $X$ by a calibration curve which is a straight line.
- The ability of inverse-predicted concentrations $Z_i$ to be proportional to the corresponding theoretical concentration $X_i$.

The first definition is applicable to calibration but has no great interest for validation as far as it is possible to accurately quantify an analyte with a nonlinear calibration curve and many novel methods are using such models. The second definition is rather a question of trueness (see Section 4.1.3). It is the more interesting definition applicable to the validation of a method. However, many guides still insist on checking the linearity of the calibration model. In other words, numerous guidelines consider that the function applicable to all the relations between the analytical responses and the concentrations to be used for the calibration must be able to follow a linear relationship. This is wrong, and the absence of calibration curve linearity does not mean that the method cannot accurately quantify, as demonstrated by the ELISA example in the previous chapter. If a linear calibration curve is strictly required, but a curvature is observable, there are two solutions:

- Limit the calibration range to the strictly linear part. This limit will become a boundary beyond which quantification is forbidden. If an instrumental response is obtained for a sample that exceeds the bounds of the calibration range, it is

mandatory to dilute the preparation to remain within the prescribed zone. This is very restrictive, whereas the calculation of a quadratic model is quite simple.
– The second solution is to estimate the impact of this curvature on the predicted concentrations. If the bias on calculated concentration is limited, the calibration model, even if not perfectly adjusted, could be further considered.

Curvature is frequent at high concentrations, but it can also be observed in the vicinity of the blank. According to our experience, this second type of curvature is mostly found in chromatography or with various detectors, including mass spectrometers. The exact positioning of the linearity limit is tricky although there is an iterative method of finding it. The simplest procedure described in the literature is to proceed empirically using a lack-of-fit test [8]. This implies a preliminary visual examination of the calibration curve. Therefore, the proposed method requires a critical mind for its interpretation. In any case, it is essential, as a prerequisite, to carry out a graphic presentation of the results to evaluate whether a test is necessary or not.

### 2.4.3 Statistical Aspects of Standard Addition Method (SAM)

In Section 1.4.1, the principles of the simple standard addition method (SAM) were presented, and the accompanying calculation was explained. Another method consists of extrapolating the with-matrix straight line to a null response, as illustrated on Figure 2.5, using unpublished personal data. At the crossing point, it is possible to compute the extrapolated concentration, noted $Z^*$ and equal to 439 ng/ml. The Resource D worksheet called SAM (Excel) illustrates the procedure when there are several standard additions.

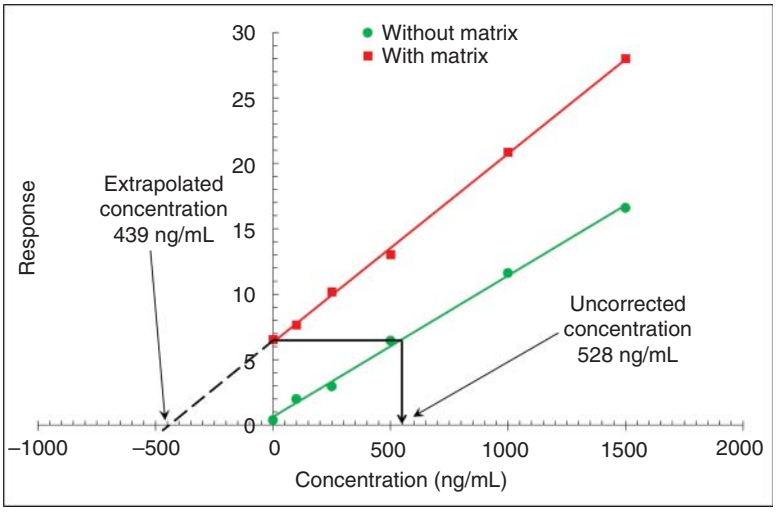



**Figure 2.5** SAM – multiple point standard addition method.

This more complex design with six standard additions is recommended by some official bodies when it is necessary to demonstrate that there is no matrix effect or when using a surrogate matrix [9]. In this example, six points are measured with and without the matrix, making altogether 12 measurements. It is possible to compute with a worksheet the coefficients of the two straight lines by using the `LINEST` built-in function. Some complementary notations will help understand the rationale.

Calibration line without matrix

$$Y = a_0 + a_1 X + e$$

Calibration line with authentic matrix

$$Y = b_0 + b_1 X + e$$

Extrapolated concentration

$$Z^* = \frac{b_0}{b_1}$$

Uncorrected concentration

$$Z = \frac{b_0 - a_0}{a_1}$$

Recovery yield

$$RY\% = \frac{Z^*}{Z} \times 100$$

One goal of SAM is to obtain the recovery yield relevant to correcting the result of the unknown sample in presence of matrix effect, or in absence of blank matrix that can be spiked to build a representative external calibration curve. However, the procedure is also useful during the method development to check the method trueness. For this reason, some guidelines recommend checking if both straight lines are parallel, as explained in Section 1.3, in other words if the difference between slopes is statistically equal to zero. In this specific context, an easy procedure to compare slope coefficients consists in simply comparing the confidence intervals of the slopes as explained in Section 2.2.

This simplified method is acceptable because both lines are assumed to have comparable statistical properties; otherwise, more classic statistical tests must be applied. If the confidence intervals overlap, this means that the difference is statistically not significant. In the example of Resource D, the 95% confidence intervals of $a_1$ and $b_1$ are:

| Presence of an authentic matrix | 95% slope confidence interval |
|---|---|
| No | [0.0100; 0.0116] |
| Yes | [0.0137; 0.0151] |

Because intervals are not overlapping, it can be concluded that the two slopes are significantly different, with a risk of 5%, and the lines of Figure 2.5 are not parallel.

This sample shows a matrix effect. It is also possible to propose a graphical application of the method. An alternative confirmation consists in verifying if relative difference between slope coefficients is included in an acceptance interval, for instance, of ±15% as defined by the European Medicines Agency [10]. If we consider that the line without matrix to be the reference, the relative difference between the two slopes is 33.7%:

$$\frac{|a_1 - b_1|}{a_1} = \frac{|0.0108 - 0.0144|}{0.0108} = 0.337 \text{ or } 33.7\%$$

This value is far from 15% but not surprising because the two lines are obviously non-parallel and it can be concluded that the matrix effect is significant. Consequently, the result must be corrected using the recovery yield to define a correction factor. For this example, the recovery yield is 82.7% and the correction factor is 0.832. For instance, it can be applied to the uncorrected concentration that gives $527 \times 0.832 = 439$. This correction factor can be applied to any unknown sample as far as the matrix is analogous. Result correction is of some consequence on the MU and a full example is presented in Section 8.4.2 showing how to take account of the correction factor in estimating MU.

**Resource D Standard addition method (Excel).**

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Resource D: Standard addition method** | | | |
| 2 | | Response | | |
| 3 | Concentration (ng/ml) | Without matrix | With matrix | |
| 4 | 0 | 0.37 | 6.54 | |
| 5 | 100 | 1.99 | 7.64 | |
| 6 | 250 | 2.95 | 10.16 | |
| 7 | 500 | 6.43 | 13.03 | |
| 8 | 1000 | 11.60 | 20.85 | |
| 9 | 1500 | 16.59 | 27.99 | |
| 10 | **Results** | | | |
| 11 | Slope | Intercept | | |
| 12 | Without matrix | | | |
| 13 | 0.0108 | 0.6382 | =LINEST(B4:B9;A4:A9;TRUE;TRUE) | |
| 14 | 2.805E-04 | 0.2165 | | |
| 15 | 0.997 | 0.366 | | |
| 16 | 1476 | 4 | | |
| 17 | 198 | 0.536 | | |
| 18 | With matrix | | | |
| 19 | 0.0144 | 6.3250 | =LINEST(C4:C9;A4:A9;TRUE;TRUE) | |
| 20 | 2.368E-04 | 0.182734888 | | |
| 21 | 0.999 | 0.308959382 | | |
| 22 | 3700 | 4 | | |
| 23 | 353 | 0.382 | | |
| 24 | | | | |
| 25 | Uncorrected concentration | 527.7 | =(B19-B13)/A13 | |
| 26 | Extrapolated concentration | 439.1 | =B19/A19 | |
| 27 | Recovery yield | 83.2% | =B26/B25 | |
| 28 | **Confidence intervals of slopes** | | | |
| 29 | Risk of error | 0.05 | | |
| 30 | Student ' s t | 2.776 | | =T.INV(1-B29/2;B16) |
| 31 | Without matrix | 0.0100 | 0.0116 | =A13-B30*A14 |
| 32 | With matrix | 0.0137 | 0.0151 | =A19-B30*A20 |

The spiking sample matrix is sometimes misunderstood by analysts, and the main criticism is that the analyte, added as a surrogate calibrant, is not in the same chemical form as the authentic analyte present in the sample. We already have shown in Chapter 1 that in many situations, the compound used for calibration is identical but many solutions exist to get around this problem. May this criticism be valid? Experience shows that, generally, the recovery yield thus obtained is operational and does not impede a good decision.

As already explained, the use of stable isotope labeled (SIL) calibrant or spiking compound may prevent this bias and bring the method closer to a primary method of analysis. This is the recommendation made by official guidelines for the validation of analytical methods for chemical pollutants, such as pesticides, dioxins, polycyclic aromatic hydrocarbon (PAH), or acrylamide. It can be noted that in recent years, the catalogs of commercial reagents have been greatly enriched with this type of molecule.

It has been suggested that when using a single spiked sample, it should have a concentration at least five times that of the test sample, but linearity must be checked before. If the response of the system is nonlinear, the extrapolation involved in the SAM approach becomes more problematic.

There is another extension of SAM in a multidimensional way, called H-point standard addition method (HPSAM), illustrated with an example of drug control in Section 10.1.

## 2.5 Metrological Approach to Calibration

From a metrological perspective, calibration is defined in International Vocabulary of Metrology (VIM) as:

> "An operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication" [11].

This definition of this operation is clearly a two-step procedure as already underlined. If the so-called "relation" is noted as already proposed $Y = f(X) + E$, $X$ is the "measurement standard" value and $Y$ is the "indication". The $f$ function establishes the relationship between $X$ and $Y$. As discussed in Chapter 1, $Y$ may be diverse: a simple electrical signal, a peak area, a peak height, a ratio between two peaks, etc. This dual nature of calibration may be more precisely established with the mathematical notation:

Step 1. Calibration

$$Y = \alpha_0 + \alpha_1 X + E \tag{2.38}$$

Predicted response for known $X$

$$Y = a_0 + a_1 X \tag{2.39}$$

Step 2. Inverse predicted concentration

$$Z = \frac{Y - a_0}{a_1} \tag{2.40}$$

The use of Greek letters in Eq. (2.38) and Latin in Eq. (2.39), is aiming to highlight that coefficients $a_0$ and $a_1$ are simply estimates of $\alpha_0$ and $\alpha_1$, and may vary according to the number of calibrators, their distribution across the calibration range, the number of replicates, the estimation technique, etc. as discussed further. Several approaches are available in the literature to calculate these coefficients, but the most common is OLS regression presented in Section 2.3.1. Because the VIM definition of calibration does not make the difference between the authentic analyte and the calibration compound, which can be authentic or surrogate, some decades ago, several authors proposed to directly fit the reverse function that links $X$ to $Y$, giving for the linear case:

Step 1. Reverse calibration

$$X = \beta'_0 + \beta'_1 Y + E' \tag{2.41}$$

Step 2. Predicted concentration

$$Z = a'_0 + a'_1 Y \tag{2.42}$$

This approach raises different questions. Originally, the major assumption when using OLS regression is that $X$ is the independent predictive variable known to be without any error and $Y$ is the dependent predicted random variable. In Eq. (2.3), $E$ is the error in $Y$ while the error in $X$ is assumed to be zero or, at least, negligible. Contrarily, in Eq. (2.41), $E'$ is the error in $X$ while there is no error in $Y$.

Because $X$ and $Y$ do not play the same role in the OLS, as it aims to minimize the sum of squared residuals in the $Y$-dimension, on the other hand, reverse calibration uses the same criterion, but on $X$-dimension. Thus, the impact of calibration on the MU is different. There is also a direct link between the values of coefficients $a$ and reverse coefficient $a'$ depending on the coefficient of determination $r^2$ exhibited by the following equations:

Inverse and reverse calibration links between $a$ and $a$'coefficients

$$\begin{cases} a'_0 = \dfrac{(1 - r^2)\overline{Y} - a_0}{b_1} \\ a'_1 = \dfrac{r^2}{a_1} \end{cases} \tag{2.43}$$

Equation (2.43) show that both approaches give close results if $r^2$ is close to 1. This condition is always justified for calibration. A review published some years ago explained the statistical problems raised by both possible approaches [12]. The major drawback is that reverse calibration is relatively easy to apply when the model is a straight line and more difficult for others. While more recent methods use nonlinear calibration models or even require weighed algorithms, reverse calibration did not become popular among analysts except for a few analytical techniques.

The most famous multidimensional application of reverse calibration is near-infrared spectrometry (NIRS), which is widely used to determine proximate components in foods or monitor certain pharmaceutical molecules. In this context, since it is impossible to prepare standard solutions containing pure analytes, calibration is obtained by fitting near-infrared spectra to measurements obtained on the same sample by another reference method. As calibration is performed using full spectra containing many wavelengths it is called *multivariate calibration*. The model becomes:

$$\mathbf{z} = g(\mathbf{Y}) + \mathbf{E}$$

$\mathbf{z}$ is the vector of measurements obtained by the reference method, $\mathbf{Y}$ and $\mathbf{E}$ are matrices – in the mathematical sense – i.e. sets of instrumental responses and errors, traditionally denoted in bold type. This calibration procedure requires specific statistical processing. The best known is the partial least-squares regression (PLS) described in specialized books and out of the scope of this book [13, 14].

### 2.5.1 Errors in Inverse-predicted Values

As explained in Section 6.6, second step in MU estimation procedure consists in obtaining estimates of the standard deviations of the different identified uncertainty sources. When inverse predicting $Z$ the sample concentration, several sources of uncertainty are included within the instrumental response, such as sample preparation or instrument instability. Whereas $Z$ is a random variable, it is appealing to compute the standard deviation $s_Z$ and use it for MU estimation.

There is some paradox in doing that, while it must be reminded that $Z$ is comparable to $X$, which is supposed to be error-free to comply with the assumptions of the OLS algorithm. Consequently, $Z$ should also be error-free. Despite this contradiction, an estimate of the standard deviation $s_Z$ of the inverse-predicted concentration was developed and given in Eq. (2.44) [8].

Standard deviation of inverse-predicted concentration $Z$

$$s_Z = \frac{s_E}{a_1}\sqrt{\frac{1}{J} + \frac{1}{I} + \frac{(Y_k - \overline{Y})^2}{a_1^2 \sum_i (X_i - \overline{X})^2}} \tag{2.44}$$

where:

- $s_E$ is the residual standard deviation of calibration given by Eq. (2.16).
- $I$ is the number of calibrators $1 \le i \le I$.
- $J$ is the number of replicate measurements of the study sample. If no replicate is present, the resulting quotient is $1/J = 1$.
- $X_i$ is the calibrant concentration.
- $Y_k$ is instrument signal of study sample.
- $\overline{X}$ and $\overline{Y}$ are the mean values of concentrations and responses of calibrators, respectively.

This equation is uneasy to use as it requires knowing, in advance, how many replicates will be done. When SAM is used, it is also possible to derive standard deviation

$s_{Z^*}$ of the extrapolated concentration $Z^*$ from Eq. (2.45). This formula will be used for the examples of Section 10.1.

Standard deviation of extrapolated concentration $Z^*$

$$s_{Z^*} = \frac{s_E}{a_1}\sqrt{\frac{1}{I} + \frac{\overline{Y}^2}{a_1^2 \sum_i (X_i - \overline{X})^2}} \tag{2.45}$$

The use of these standard deviations is not suitable for practical MU estimation and we will propose a more global and comprehensive approach in Section 7.2. But these equations can provide valuable information regarding the optimization of external calibration performances. The following rules inferred from Eq. (2.44) can be applied to better plan the calibration experiment and define some "best calibration practices":

- *Rule 1*. The first equation term includes the residual standard deviation $s_E$ which is an estimate of the random errors of instrumental responses. The possibility of minimizing this parameter is provided by using an appropriate internal standard, as recommended in Section 1.3.
- *Rule 2*. The second term $a_1$ matches with the slope of the calibration curve. The higher the slope is, the lower the errors in the inverse-computed concentration $Z$. As explained in [15], $a_1$ is understandable as the sensitivity of the analytical method, the latter being defined as the ratio between the response variation of the analytical calibration and the analyte quantity variation. An analytical method can thus be considered sensitive when a small variation in the calibrant concentration induces a large variation in the response. This definition differs from international guidelines such Food and Drug Administration (FDA) and European Medicines Agency (EMA), where sensitivity is described as the "lowest measurement range with acceptable accuracy and precision." It must be noted that when the analytical response is corrected by an internal standard, it represents the slope of the ratio and not the authentic analyte response.
- *Rule 3*. At least two strategies for calibrator selection and distribution can be considered to reduce the size of the $Z$ standard deviation and consequently its confidence interval. First, the number $I$ of calibration standards can be increased. According to some guidelines, approximately six calibration points are adequate in many experiments. This point is open to discussion in an integrated approach. The second strategy relies on increasing the number of replicates $J$, which also reduces the width of the confidence interval.
- *Rule 4*. When the working sample instrumental response $Y_k$ is close to the average point of the calibration range, i.e. approaches mean $\overline{Y}$, the third term inside the square root converges to zero, thus reducing the $s_Z$ value. In the conventional OLS approach, the most precise results are obtained when the measured signal corresponds to a point near the average point (centroid) of the regression line. Obviously, prediction errors are not equal for all points and are smaller when the response is close to the centroid rather than at the edges. This drawback is enhanced when the calibration range is large and response $Y$ becomes heteroscedastic as explained in Section 2.3.2 about WLS regression.

– *Rule 5.* When looking at the quantity $\sum_i (X_i - \overline{X})^2$ it can be deduced that distributing calibrators at the ends of the calibration range is more valuable since it is maximized. One of the well-known properties of least-squares technique is that better estimations of the regression coefficients are obtained when two levels of concentration are set far apart in the calibration interval. For instance, international guidelines to analytical validation, such as FDA and EMA recommend performing a calibration with at least six calibrators, two of them located at expected upper and lower limits of quantification (ULOQ and LLOQ) in the calibration design. However, no other calibrator locations are indicated.

Official guidelines have usually recommended most of the rules listed above. This leads to satisfactory inverse-prediction when dealing with External Calibration (EC) in the linear case. While only a few formal expectations are provided for other types or models of analytical calibration and the multipoint matrix-matched external calibration (MMEC) is recommended by international guidelines for the validation of bioanalytical methods [9]. This mention of official guidelines suggests a growing interest in the need for new analytical calibration practices. Currently, the possibility of cross validating the results obtained with an alternative analytical method to a reference one allows the analyst to investigate alternative quantification methodologies without sacrificing performance.

"Obligations of result" and "obligations of means" have a longstanding history in private law. The obligation of result is simply the obligation of the debtor to attain a predetermined result. These legal concepts could be transposed to the organization of the method validation. For instance, official recommendations are often based on an "obligation of means" such as a minimal number of calibrators or replicates for calibration conditions. If an "obligation of results" were prescribed, such as participating in systematic proficiency testing (Section 4.3), the analyst would be able to select the calibration organization most appropriate to the laboratory.

The implementation of this other approach may promote renewed criteria to evaluate the calibration performance, including a focus on the observed results and their respective uncertainties. The comprehensive validation approach is probably the better way to establish innovative quantification methodologies for modern instrumentation. It also means that calibration is a fully integrated part of the analytical process.

### 2.5.2 Calibration as a Source of Uncertainty

When considering MU estimation of a quantitative result, the calibration step remains one of the key stages to reliably assess its quantification. Unfortunately, it seems impossible to estimate the specific role of calibration in MU. Only a comparative approach is applicable using different calibration models or algorithms. Since the 1990s, metrologists have developed what is called the Uncertainty Approach in opposition to the more traditional Error Approach. To estimate the MU of any reported measurement value, an effective solution consists of applying the

uncertainty approach and considering calibration as one of the many sources of uncertainty participating in the combined MU of an analytical result, noted $u_c(Z)$ with subscript $c$ stands for combined.

As recalled above, many studies have investigated the influence of different experimental designs and strategies applicable to calibration, such as WLS regression, replication of measurements or optimized distribution of calibrators within the calibration interval. If they are beneficial, they should be evaluated due to the reduction of $u_c(Z)$.

The estimation of MU in analytical sciences is still at its beginnings, and no absolute procedure is recognized by analysts. Examples applicable to a specific operating procedure or field of analysis have been published. However, the role of calibration in MU must be scrutinized. The general procedure to estimate MU, is described in Chapter 6. It is relatively straightforward and consists of four steps: specify the measurand; identify the uncertainty sources; simplify and quantify the uncertainty components; and calculate the combined uncertainty. Calibration is one of the many sources of uncertainty in analytical sciences which can be identified in the second step. Whatever the calibration procedure, at least three major sources are present:

- The chosen calibration model.
- Applied algorithm to estimate model coefficients.
- In some cases, calibrant uncertainty in the reference material used for calibrator can also be accounted for.

However, it is difficult to define an experimental design to separately estimate the uncertainty linked to calibration. Therefore, the role of calibration in MU can only be indirectly estimated as already underlined. In some statistical textbooks, strategies to select the *best* calibration model and minimize the importance of calibration as an uncertainty source are proposed. Unfortunately, in analytical sciences the $X$ concentration and $Y$ response are naturally highly correlated and classical statistical criteria, such as the $r^2$ or $AIC$, are inefficient for selecting the most appropriate calibration model. Several other techniques exist, such as the visual inspection of residuals or dedicated statistical testing to verify the lack-of-fit or the homoscedasticity.

But experience shows that all these techniques are not very efficient because of the extremely strong causal relationship between $X$ and $Y$. The Guide to the expression of Uncertainty in Measurement (GUM) general procedure recommends regrouping sources of uncertainty, and a more empirical approach consists of making several estimations of the combined MU using different calibration models and regression techniques. For example, using QC values or the method accuracy profile are possible alternatives to reach this goal. To obtain reliable quantification in terms of trueness and precision, the analytical calibration function must take the response relationships for all relevant analytes and interferences into account. From a practical perspective, two primary prerequisites must be fulfilled:

- Physicochemical properties of surrogate standard must be as close as possible to the authentic analyte in terms of solvent and nature.

– Calibration standards and working samples must have comparable behavior in the measurement system.

Finally, many MU guidelines explain how to estimate MU for one working sample. In many contract laboratories, the composition of received sample may be highly variable. The application of the proposed procedures to each working sample would be tedious and difficult. In Section 7.6.3 we promote the concept of Uncertainty Function which represents another approach more suitable to analyst's daily practice.

## References


1 Gassner, A.I., Schappler, J., Feinberg, M., and Rudaz, S. (2014). Derivation of uncertainty functions from validation studies in biological fluids: application to the analysis of caffeine and its major metabolites in human plasma samples. *Journal of Chromatography A* 1353: 121–130.

2 Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley.

3 Gort, S.M. and Hoogerbrugge, R.A. (1995). User-friendly worksheet program for calibration using weighted regression. *Chemometrics and Intelligent Laboratory Systems* 28: 193–199.

4 Robison-Cox, J.F. (1995). Multiple estimation of concentrations using logistic models. *Journal of Immunological Methods* 186: 79–88.

5 Huet, S., Bouvier, A., Poursat, M.A., and Jolivet, E. (2004). *Statistical Tools for Nonlinear Regression*, 2e. New-York: Springer-Verlag.

6 Agterbendos, J. (1979). Calibration in quantitative analysis: part 1. *Analytica Chimica Acta* 108: 315–323.

7 Agterbendos, J. (1981). Calibration in quantitative analysis: part 2. *Analytica Chimica Acta* 132: 127–137.

8 Miller, J.N., Miller, J.C., and Miller, R.D. (2018). *Statistics and Chemometrics for Analytical Chemistry*, 6e. England: Pearson Education Limited.

9 Food and Drug Administration (FDA) (2018). *Bioanalytical Method Validation Guidance for Industry*. Washington, DC: Office of Communications, Division of Drug Information Center for Drug Evaluation and Research. https://www.fda.gov/files/drugs/published/Bioanalytical-Method-Validation-Guidance-for-Industry.pdf (accessed 23 July 2023).

10 European Medicines Agency (EMA) (2023). ICH guideline M10 on bioanalytical method validation - step 5b, EMA/CHMP/ICH/172948/2019, Committee for Human Medicinal Products.

11 Bureau International des Poids et Mesures (BIPM) (2012). *International Vocabulary of Metrology — Basic and General Concepts and Associated Terms* (VIM3), JCGM 200:2012, BIPM, Sèvres, France, (aka ISO/CEI Guide 99). https://www.bipm.org/ (accessed 23 July 2023).

12 Osborne, C. (1991). Statistical calibration, a review. *International Statistical Review* 59 (3): 309–336.

**13** Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58: 109–130.
**14** Abdi, H., Chin, W.W., Vinzi, V.E. et al. (2013). *New Perspectives in Partial Least Squares and Related Methods*. Springer-Verlag.
**15** Currie, L.A. (1999). Detection and quantification limits: origins and historical overview. *Analytica Chimica Acta* 39: 127–134.

# 3

# Precision

## 3.1 Outputs of Interlaboratory Studies

### 3.1.1 Diverse Precision Parameters

Precision is one of the major characteristics of any measurement method and is defined in clause 2.5 of the VIM [1]:

> Precision is the "closeness of agreement between measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions."

The definition of precision conveys the idea that replicate measurements, when made on the same sample with the same method, are dispersed in a specific way. An early challenge for the official control bodies, which needed to compare data of different origins, was to understand how precision may vary from one laboratory to another and how they could rely on analytical data to monitor different tasks such as people's health or environmental pollution. While trade became more global, the concern of economic agents was similar for product exchanges. It was concluded that it was essential to have a common procedure for comparing measurement values obtained by different laboratories and to ascertain if the discrepancies were caused by the method or the laboratory.

Therefore, it was decided to assess precision within the framework of interlaboratory studies. Early studies were organized in the 1940s. In the 1990s, ISO published a set of five international standards under the global reference ISO 5725 [2–6]. Several parts of this major standard are under revision. The general scope was to define a harmonized procedure for the statistical interpretation of interlaboratory studies and the computation of precision parameters. The principal role was to underline that the estimation of precision depends on the condition where replicates are conducted, i.e. which sources of variations are accounted for:

– When replicates are made in the same laboratory without modifying anything between two measurements, only random effects are counted, and the precision is estimated under *repeatability condition*.

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

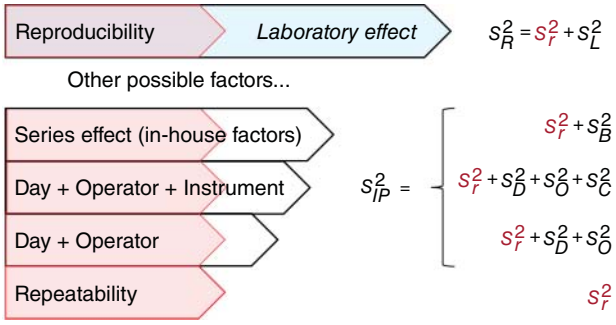



**Figure 3.1** Graphical representation of diverse total variance decomposition, affording diverse sources of variation. Repeatability is always present and represented by a red arrow.

- When replicates are performed in the same laboratory, but varying one or more operating conditions (also called factors), such as day, reagent batches, operator, instrument settings, or any other factor, the precision is estimated under *intermediate precision condition*, thus-called because it is in between repeatability (the smallest variation) and reproducibility (largest variation).
- When replicates are performed in different laboratories on the same sample with supposedly the same operating procedure, the precision is estimated under *reproducibility condition*. This means that different test portions of the working sample are analyzed by different operators on different days using almost identical standard operating procedures but adapted within different laboratories.

Though this book deals with in-house validation, the interlaboratory approach to precision is interesting because concepts and computation techniques are transposable from interlaboratory to in-house validation. For statisticians, the parameter used to estimate data dispersion is the variance or its square root, the standard deviation. Therefore, when speaking about precision, it means speaking about variances or standard deviations. Figure 3.1 is an attempt to illustrate the complexification of the precision variance from the repeatability condition, where only random effect plays a role, to the reproducibility condition, where sources of variation are numerous.

### 3.1.2 Role of Series for Data Collection

The term "series" was introduced in Chapter 2 and appears several times in all following chapters. It is critical and extensively used because it is a basic element of laboratory management and in-house validation. It can be defined as a set of measurements, repeated or not, run under repeatability condition. A series puts together all the measurements performed under the same conditions, for example, same method, same laboratory, same day, same operator, same calibration curve, etc. It is the basic planning of laboratory work that consists of grouping samples intended to be analyzed using the same analytical method. The main goal of the management of samples in series is to reduce the fixed costs of the method, as explained in Section 4.4.

When replicates are performed in a single laboratory and planned in different series, we are under intermediate precision

variation are introduced, such as days (D), operators (O), instrument settings I, and reagents. There is a full spectrum of possible combinations and possible intermediate precision standard deviations. Intermediate precision condition is recommended (or required) for in-house validation and measurement uncertainty (MU) estimation, as explained in Section 6.5.

Figure 3.1 illustrates an important idea that is developed in ISO 5725 standards. Precision variance can be decomposed into several components, each characterizing the effect of a factor of variation and a global replication condition. Finally, precision is quantified by different variances, which are combined. Depending on the organization of measurements, the intermediate precision variance $s_{IP}^2$ may contain a variable number of components.

Repeatability variance (or within-series)

$$s_r^2$$

Between-series variance

$$s_B^2$$

Intermediate precision variance

$$s_{IP}^2 = s_r^2 + s_B^2 \tag{3.1}$$

Between-laboratories variance (or laboratory effect)

$$s_L^2$$

Reproducibility variance

$$s_R^2 = s_r^2 + s_L^2 \tag{3.2}$$

An example of an interlaboratory study is used to illustrate how precision parameters can be computed and interpreted. It consists of a panel of 11 laboratories that were requested to make three replicate measurements of Pb on a sample of homogenized freeze-dried mussel tissue. The resulting dataset, called LEAD, is also used in Section 7.4 to illustrate a possible procedure for evaluating MU from the results of an interlaboratory analysis.

This dataset is described in Table 3.1. Strictly speaking, the definition of the reproducibility condition is not exactly satisfied because participants used their own analytical technique. Results for each participant are collected in Table 3.2.

**Table 3.1** LEAD – description of the dataset.

| **Title of the dataset** | **LEAD** |
|---|---|
| Reference | [7] |
| Number of laboratories | 11 participants ($I = 11$) |
| Replicates/laboratory | ($J = 3$) |
| Measurand | Lead concentration in mollusk tissue, in mg/kg |
| Methods | ICP-MS spectrophotometry |

**Table 3.2** LEAD – original dataset.

| | Replicates (mg/kg) | | | |
|---|---|---|---|---|
| **Laboratories** | **1** | **2** | **3** | **Average** |
| Lab 01 | 2.08 | 2.00 | 2.01 | 2.03 |
| Lab 02 | 2.00 | 1.93 | 1.89 | 1.94 |
| Lab 03 | 2.10 | 2.44 | 1.96 | 2.17 |
| Lab 04 | 2.45 | 2.34 | 2.49 | 2.43 |
| Lab 05 | 1.95 | 1.89 | 1.93 | 1.92 |
| Lab 06 | 1.85 | 1.91 | 1.89 | 1.88 |
| Lab 07 | 2.01 | 2.00 | 2.06 | 2.02 |
| Lab 08 | 2.00 | 2.09 | 1.98 | 2.02 |
| Lab 09 | 2.11 | 2.03 | 2.14 | 2.09 |
| Lab 10 | 2.02 | 1.98 | 1.97 | 1.99 |
| Lab 11 | 2.02 | 2.00 | 2.04 | 2.02 |

Precision parameters calculated following ISO 5725-2 recommendations are:

| **Parameters** | **Symbol** | **Value** |
|---|---|---|
| Number of laboratories | $I$ | 11 |
| Number of replicates | $J$ | 3 |
| Grand mean | $\overline{\overline{Z}}$ | 2.047 |
| Repeatability variance | $s_r^2$ | 0.00751 |
| Interlaboratory variance | $s_L^2$ | 0.01945 |
| Reproducibility variance | $s_R^2$ | 0.02696 |
| Repeatability std. dev. | $s_r$ | 0.0867 |
| Reproducibility std. dev. | $s_R$ | 0.1642 |

As stated before, the scope of ISO 5725 standards was to define common rules for comparing results issued from several laboratories and solving eventual trade disagreements. Therefore, the maximum acceptable difference between two measurements from two different laboratories is a classic acceptance criterion. It is called the Reproducibility Limit and conventionally denoted $R$. It is derived from the standard deviation of reproducibility and is equal to:

$$R = 2.83 \times s_R$$

The coefficient 2.83 is obtained by considering the 95% confidence interval of the difference between two measurements having the same variance $s_R^2$. The origin of this coefficient is linked to the variance of the difference between two random

variables with the same variance. The limit of reproducibility is obtained using the following computational rules:

Difference between lab A and lab B

$$D = |Z_A - Z_B|$$

Variance of the difference

$$s_D^2 = 2 \times s_R^2$$

Standard deviation of the difference

$$s_D = \sqrt{2} \times s_R$$

Limit of reproducibility: 95% confidence interval of the difference

$$2.0 \times s_D = 2.0 \times \sqrt{2} s_R = 2.83 \times s_R$$

The coefficient 2.0 used in the last formula is called the "coverage factor." This is standardized and explained in Section 6.2. This gives, for the LEAD example:

$$R = 2.83 \times 0.1642 \simeq 0.465 \text{ mg/kg}$$

A decision rule can be derived from this criterion as follows: if two lead measurements performed in two different laboratories on a sample around 2 mg/kg differ by less than 0.465 mg/kg, they are considered equivalent; both laboratories give the same result. Figure 3.2 illustrates the LEAD dataset. Measurements are spread around the grand mean, and some results from laboratories 3 and 4 are not highly consistent with the results from the others; several of them do not comply with the limit of reproducibility. The question of abnormal results will be addressed in Section 3.4.2 about outliers.

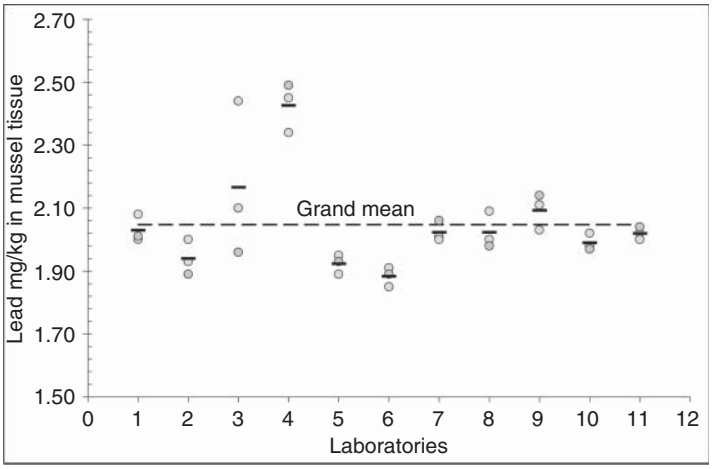



**Figure 3.2** LEAD – illustration of interlaboratory study. The grand mean is a dotted line, small horizontal bars are laboratory means.

## 3.2 Analysis of Variance (ANOVA)

### 3.2.1 Computation of Precision Parameters

The idea behind interlaboratory data interpretation is that each measurement value is a combination of a common repeatability precision that characterizes the method and the effect of the laboratory bias. In an interlaboratory study, the effect of the laboratory factor is the influence of each participant on the determination of the contents $X$ of the material sent to participants. The corresponding model is:

$$Z_{ij} = X + L_i + E_{ij}$$

- $Z_{ij}$ is the measurement value of replicate $j$ with $(1 \leq j \leq J)$ for laboratory $i$ with $(1 \leq i \leq I)$.
- $X$ is the "true value" of the study material (a discussion about this concept is proposed in Section 4.1.2). It is constant.
- $L_i$ is the effect of the $i$th laboratory on the global result. It is not constant and may take random values.
- $E_{ij}$ is the random residual error of each replicate.

Therefore, total precision, i.e. total variance of $Z_{ij}$ must be decomposed into two main components: the laboratory effects and the residual errors. The classic statistical tool used to decompose the variance is the analysis of variance (ANOVA). It was first described in the 1920s by Fisher. Its principle is to distribute the total variance of a dataset into several pieces, each corresponding to a source of variation. This operation is possible if the dataset has a predefined structure.

For instance, the LEAD dataset structure is simple: each data is a replicate, hierarchically linked to one laboratory. The laboratory entity is called a *factor*. Many types of ANOVA may exist, depending on the structure of the dataset and the number of factors. According to the nomenclature adopted by statisticians, the ANOVA used for interpreting interlaboratory study is a *one-way random effects ANOVA* as opposed to a *one-way fixed effects ANOVA*.

The effects $L_i$ of the laboratory factor are said to be random because they are not controlled by the organizer, who randomly selects a group of laboratories among all existing ones [8, 9]. On the contrary, fixed factor levels are controlled, like temperature, volume, or pH. Random and fixed factors can be mixed into more complex designs that require specialized software to be processed. An example of a more complex variance decomposition model is available in [10].

Any participating laboratory must perform replicates because the goal is to measure variances. For each laboratory, it is essential that replicates be performed under repeatability conditions, i.e. within the same series. Measurement replication is no longer required if the laboratory takes part in proficiency testing. This is another type of collaborative study devoted to estimating trueness but not precision. The statistical modeling of proficiency testing is different, as will be explained in Section 7.4.

In the context of in-house validation, the factor used to do the ANOVA is no longer the laboratory but the series, which often corresponds to the date when the assays

were completed, but also when instrument settings were revised, or new reagents were prepared, etc. The mathematical model is the same, except that the effect for series is denoted $B_i$ instead of $L_i$:

$$Z_{ij} = X + B_i + E_{ij}$$

Though the most classic application of this modeling is described in ISO 5725 standards for an interlaboratory study, the same mathematical approach is fully valid when applied to measures done in-house and under intermediate precision conditions. In both cases, data are structured according to a single factor, giving the name one-way ANOVA. As stated, the total variance is decomposed into two variances, one corresponding to the factor effects and the other to the errors, respectively. With more complex data structures, the decomposition must consider this complexity.

The starting point of the decomposition is quite simple, as illustrated by the basic equation (3.3). The total mean $\overline{\overline{Z}}$ is computed with the whole dataset and conventionally named the "grand mean."

Basic equation

$$Z_{ij} - \overline{\overline{Z}} = (Z_{ij} - \overline{Z_i}) + (\overline{Z_i} - \overline{\overline{Z}}) \tag{3.3}$$

*i*th laboratory mean

$$\overline{Z_i} = \frac{\sum_{j=1}^{J} Z_{ij}}{I} \tag{3.4}$$

Grand mean

$$\overline{\overline{Z}} = \frac{\sum_{i=1}^{I} \overline{Z_i}}{I} \tag{3.5}$$

Squared basic equation

$$(Z_{ij} - \overline{\overline{Z}})^2 = [(Z_{ij} - \overline{Z_i}) + (\overline{Z_i} - \overline{\overline{Z}})]^2$$

Developed squared basic equation

$$\sum_i \sum_j (Z_{ij} - \overline{\overline{Z}})^2 = \sum_i \sum_j (Z_{ij} - \overline{Z_i})^2 + \sum_i (\overline{Z_i} - \overline{\overline{Z}})^2 2 \sum_i \sum_j (Z_{ij} - \overline{Z_i}) \times (\overline{Z_i} - \overline{\overline{Z}}) \tag{3.6}$$

The basic equation is a simple set of additions and subtractions of different deviations. It may be surprising to square this equation, but if we try to compute the mean value of each difference, it is easy to see that all sums are equal to 0. Moreover, it can be demonstrated that the double cross-product of Eq. (3.6) is always equal to 0.

$$2 \sum_i \sum_j (Z_{ij} - \overline{Z_i}) \times (\overline{Z_i} - \overline{\overline{Z}}) = 0$$

The final Eq. (3.7) consists of three sums of squared deviations that can be related to the elements of the model, namely the total sum of squares, the repeatability or residual sum of squares, and the interlaboratory sum of squares.

Simplification

$$\sum_i \sum_j (Z_{ij} - \overline{\overline{Z}})^2 = \sum_i \sum_j (Z_{ij} - \overline{Z_i})^2 + \sum_i (\overline{Z_i} - \overline{\overline{Z}})^2 \tag{3.7}$$

Total sum of squares = Repeatability sum of squares +
Interlaboratory sum of squares

$$\begin{aligned} &\text{Simplified notation for in} - \text{house validation } SS_t = SS_W + SS_B \\ &\text{Simplified notation for interlaboratory study } SS_t = SS_r + SS_L \end{aligned} \tag{3.8}$$

Equation (3.7) can be written in an equivalent abridged notation, which will be used below. The following conventions are used:

- Subscript *W* means *Within*
- Subscript *B* means *Between*-series in the context of in-house validation, and *L* means *Between*-laboratories in the context of interlaboratory.
- The notation SS for sum of squares will be very useful when developing an Excel worksheet, as explained in Section 3.4.

Finally, $SS_W$ matches with repeatability. In the rest of the text, both notations are applied, and the same algorithm is used to compute precision parameters, either in the context of interlaboratory or in-house validation. ANOVA is nothing more than the application of Pythagoras's Theorem, where the square of the hypotenuse of a right-angled triangle equals the sum of squares of the other two sides. In other words, the total sum of squares is equal to the sum of squares due to repeatability and the sum of squares due to factor effects.

Figure 3.3 proposes a geometric illustration of this decomposition. Finally, the total sum of the squares of the deviations from the grand mean is decomposed into a sum of squares of deviations due to the laboratory (or series) effect and a residual sum of squares due to the repeatability error. More complete notation is required to accomplish the full calculation. Let us note $SS_i$ the sum of the squares of the deviations of laboratory *i* from its own mean $\overline{Z_i}$.

Sum of squares of laboratory *i*

$$SS_i = \sum_j (Z_{ij} - \overline{Z_i})^2 \tag{3.9}$$

Within (series or laboratory) sum of squares or Repeatability sum of squares

$$SS_r = SS_W = \sum_i SS_i \tag{3.10}$$

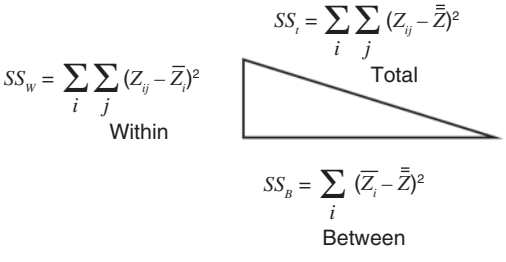



**Figure 3.3** Geometric interpretation of the general ANOVA.

Total sum of squares

$$SS_t = \sum_i \sum_j (Z_{ij} - \overline{\overline{Z}})^2 \tag{3.11}$$

$$SS_t = SS_W + SS_B \tag{3.12}$$

Between-series sum of squares or Between-laboratories

$$SS_B = SS_L = SS_t - SS_W \tag{3.13}$$

The calculation by difference of the between-series (or -laboratories) sum of squares, denoted $SS_B$ or $SS_L$ is the simplest way to avoid any further calculation. It was proposed when modern computational means were not available. Therefore, in the first publication of ISO 5725 standards in 1994, this was presented as the standardized computation method.

There is a major downside when the effects of the factor (series or laboratories) are not significantly different, i.e. when the interlaboratory or between-series variance is close to 0. By calculation, the sum of squares and the corresponding variance may be negative, which is nonsense since, by construction, a variance must always be positive! As explained below, the recommendation is to set $SS_L$ to 0. But other algorithms exist (Section 4.3.4).

Next, to calculate the variances, the *SS* are divided by the appropriate numbers of degrees of freedom. The degrees of freedom rely on an assumption about the distribution of residual errors $E_{ij}$. They are assumed to be all distributed according to identical Normal laws $\mathcal{N}(0, \sigma^2)$. This requirement is the same as for the residuals for the ordinary least-squares (OLS) method described in Section 2.3.1. If there is a way to circumvent this requirement by using weighted least-squares (WLS), there is not such a possibility with ANOVA. For instance, in Figure 3.2, most variances are identical and fulfill the basic assumption, except in Lab 03. Several statistical tests are proposed to verify this hypothesis in ISO 5725 standard. In the framework of in-house validation, we consider these tests to be useless.

The observed model in Figure 3.4a shows that the data from each laboratory are diversely scattered. This means they correspond to diverse distribution laws, each with a specific mean $\overline{Z_i}$ and standard deviation $s_i$. The size of the dotted circle illustrates the magnitude of the dispersion. When moving to the theoretical model developed by Fisher, it is assumed that all laboratories have the same standard deviation, corresponding to the repeatability standard deviation.

Figure 3.4b, illustrates the consequences of this assumption about variances in the modeling of data. In some way, the standard deviation of repeatability represents an average residual error standard deviation. At first glance, it may seem astonishing to enforce a single standard deviation and no longer consider the observed standard deviation of each laboratory. Behind this assumption stands the idea that the measuring procedure can be characterized by unique precision parameters, whoever applies it.

The differences between each laboratory mean $\overline{Z_i}$ and the reference value $X$ assigned to the test sample are called the laboratory effects and are visually

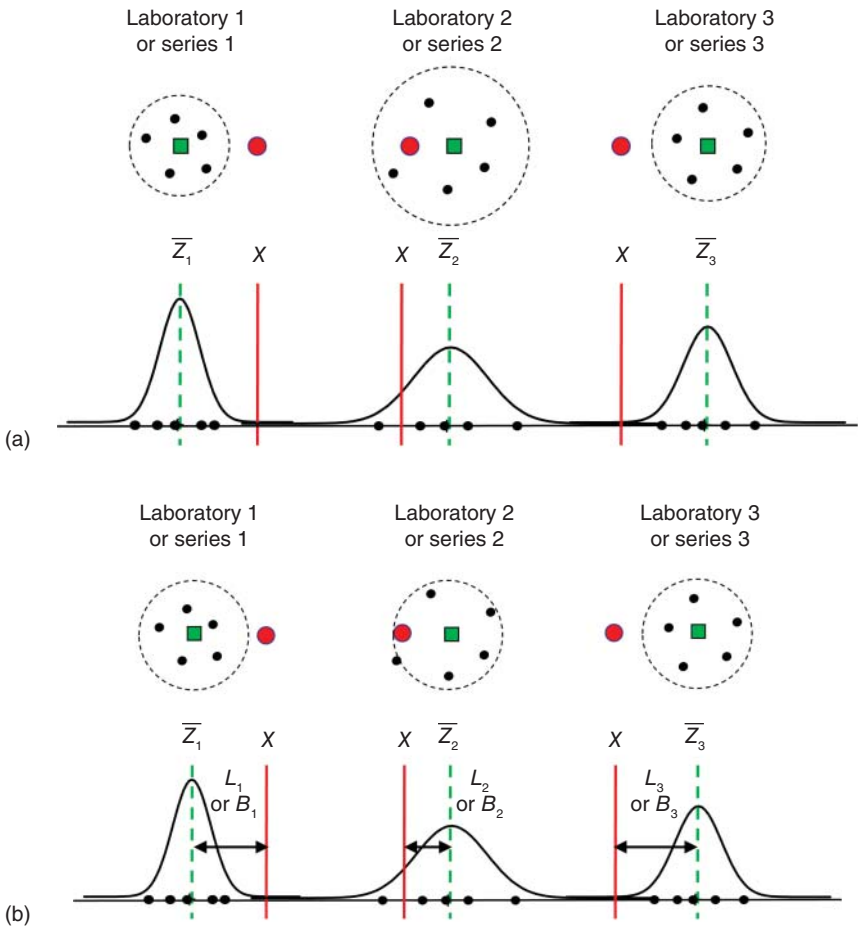



**Figure 3.4** (a) ANOVA – observed model. (b) ANOVA – theoretical model.

illustrated by double horizontal arrows. They correspond to the random variables $L_i$ (or $B_i$ for the series) of the model. For one laboratory, it is essential to understand that the laboratory effect is not constant and will also randomly vary from one replicate to another.

Finally, the formulas applicable to obtain the precision variances and standard deviations are the following. They are presented in different equivalent notations. The same formulas are valid when data are obtained under reproducibility conditions or intermediate precision conditions, although trials are not done in the same experimental framework, several laboratories, or a single laboratory.

For clarity, the same notation $I$ is used for different quantities (and names): for interlaboratory studies, it is the number of laboratories, while for in-house studies, it is the number of series.

Within-series variance or Repeatability variance

$$\begin{cases} s_r^2 = s_W^2 = \frac{SS_W}{N-I} \\ s_r^2 = \frac{SS_W}{I(J-1)} \\ s_r^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(Z_{ij}-\overline{Z}_i)^2}{I(J-1)} \end{cases} \quad (3.14) \quad (3.15)$$

Between-series variance or Between-laboratories variance (equivalent notations)

$$\begin{cases} s_L^2 = s_B^2 = \frac{\left(\frac{SS_B}{I-1} - s_r^2\right)}{J} \\ s_B^2 = \frac{\left(s_{\overline{Z}}^2 - s_r^2\right)}{J} \\ s_B^2 = \frac{\frac{\sum_{i=1}^{I}(\overline{Z}_i-\overline{\overline{Z}})^2}{I-1} - s_r^2}{J} \end{cases} \quad (3.16)$$

Reproducibility variance

$$s_R^2 = s_r^2 + s_L^2 \quad (3.17)$$

Intermediate precision variance

$$s_{IP}^2 = s_r^2 + s_B^2 \quad (3.18)$$

Reproducibility standard deviation

$$s_R = \sqrt{s_r^2 + s_L^2} \quad (3.19)$$

Intermediate precision standard deviation

$$s_{IP} = \sqrt{s_r^2 + s_B^2} \quad (3.20)$$

## 3.2.2 Additional Parameters

### 3.2.2.1 Relative Standard Deviation of Parameters

A long-established parameter used for precision notation, proposed in many regulatory documents and analytical standards, consists in computing the relative standard deviation (*RSD*). This parameter, sometimes called the "coefficient of variation" or *CV*, is easily obtained as the ratio of a standard deviation and the corresponding average value. For instance, for one laboratory, the *RSD* is

$$RSD_i = \frac{s_i}{\overline{Z}_i} \times 100$$

This is a dimensionless parameter, generally expressed as a percentage. When estimating the precision, different *RSD* can be computed using the corresponding standard deviation, e.g. repeatability or intermediate precision, and the grand mean. According to our notations, the *RSD* of precision may be expressed at least in three different forms.

*RSD* of repeatability

$$RSD_r = \frac{s_r}{\overline{\overline{Z}}} \times 100$$

*RSD* of intermediate precision

$$RSD_{IP} = \frac{s_{IP}}{\overline{\overline{Z}}} \times 100$$

*RSD* of reproducibility

$$RSD_R = \frac{s_R}{\overline{\overline{Z}}} \times 100$$

For instance, the *RSD* of repeatability and reproducibility obtained from the LEAD dataset are:

$$RSD_r = 100 \times \frac{s_r}{\overline{\overline{Z}}} = 100 \times \frac{0.0867}{2.047} = 4.23\%$$

$$RSD_R = 100 \times \frac{s_R}{\overline{\overline{Z}}} = 100 \times \frac{0.1642}{2.047} = 8.02\%$$

All these notations are not equivalent but are considered convenient to characterize the precision of a method because they can be used to compare different methods, whatever the concentration range. Therefore, they are intensively used to establish regulatory validation criteria, as explained in Section 7.5.1. For example, an acceptance criterion for reproducibility is based on Horwitz's model, which was developed to decide whether an observed $RSD_R$ value is acceptable or not.

It must be noted that computed *RSD* results from the ratio of two random variables, giving a new parameter with high uncertainty. In the case of method validation procedures, measurements are preferably done on special validation material, the assigned or reference value of which is known. It is then possible that the *RSD* denominator should be replaced by this value. The new ratios computed in this way are no more *RSD*, in the classical definition. They are sometimes considered as estimates of precision, regardless of the estimated trueness [11].

#### 3.2.2.2 Variance of the Grand Mean

The method accuracy profile (MAP) is a method validation procedure presented in Chapter 5. All calculations are based on data collected under intermediate precision conditions and the former equations are adequate to compute the grand mean, standard deviations of repeatability, and intermediate precision. But it also requires an estimate of the variance of the grand mean. This additional parameter is given by Eq. (3.21). It depends on the variance ratio, noted $A$ (see Eq. 3.24), which plays a fundamental role in the MAP procedure, as explained further.

Variance of the grand mean

$$s^2_{\overline{\overline{Z}}} = s^2_{IP} \times \frac{1}{IJQ} \tag{3.21}$$

Standard deviation of the grand mean

$$s_{\overline{\overline{Z}}} = s_{IP}\sqrt{\frac{1}{IJQ}} \tag{3.22}$$

Coefficient $Q$

$$Q = \frac{A+1}{J \times A + 1} \tag{3.23}$$

Variance ratio

$$A = \frac{s_B^2}{s_r^2} \tag{3.24}$$

In Section 7.2, a procedure is also proposed to derive the MU from data collected under intermediate precision condition. This is recognized as a very promising solution to obtain MU.

## 3.3 Balanced and Unbalanced Experimental Design

According to the ANOVA vocabulary, the experimental design is said *to be balanced* when the number of replicates is the same at all levels of the factor (i.e. laboratories or series). For instance, the LEAD experimental design is balanced because all participants did the same number of replicates $J = 3$. The Figure 3.5 provides a convenient graphical representation of balanced and unbalanced experimental designs. The main rectangle represents the whole dataset, and small rectangles represent laboratories or measurement series. The height is proportional to the number of laboratories $I$ and the width to the maximum number of replicates $J$. If the design is balanced, all gray rectangles are the same length. If the design is unbalanced, the length varies with the number of replicates $n_i$.

The statistical data processing described in Section 3.2 is only applicable to a balanced design. For unbalanced design, when all levels do not contain the same number of replicates, it is necessary to introduce $n_i$ the number of replicates of laboratory/series $i$. The indexing of replicates is as follows: for the balanced design, it is $1 \leq j \leq J$, and for the unbalanced design, it is $1 \leq j \leq n_i$. Previous equations are consequently modified and become:

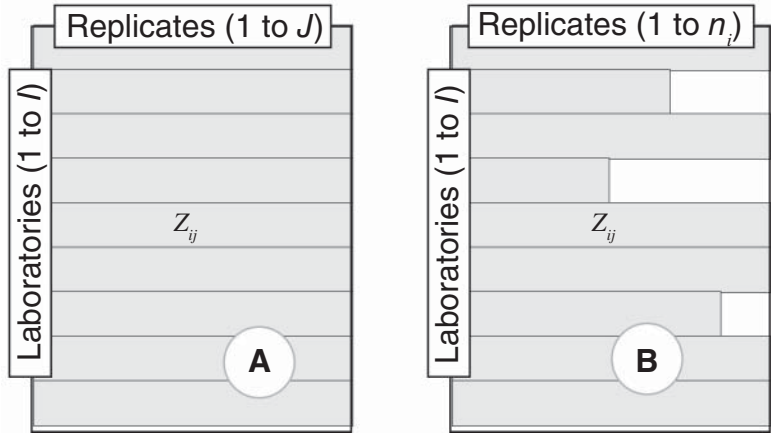



**Figure 3.5** Graphical representation of the experimental design. (a) Balanced and (b) unbalanced.

Total number of measurements

$$N = \sum_i n_i \tag{3.25}$$

Corrected number of measurements

$$N' = N - \frac{\sum_i n_i^2}{N} \tag{3.26}$$

General ANOVA equation

$$\sum_i \sum_j (Z_{ij} - \overline{\overline{Z}})^2 = \sum_i \sum_j (Z_{ij} - \overline{Z_i})^2 + \sum_i n_i (\overline{Z_i} - \overline{\overline{Z}})^2 \tag{3.27}$$

$$SS_t = SS_W + SS_B$$

Repeatability variance

$$s_r^2 = \frac{SS_W}{N - 1} \tag{3.28}$$

Between-series variance or interlaboratory variance

$$s_B^2 = \frac{(I - 1)\left(\frac{SS_B}{i-1} - s_r^2\right)}{N'} \tag{3.29}$$

These modifications are not of utmost importance, but interesting for software development. Obviously, when *RSD* are computed, they are modified in the same way as the standard deviation and grand mean are. Finally, the estimation of the precision parameters is more sophisticated than sometimes presented. Quite often, analysts claim they have obtained "a repeatability" by simply computing the average value $\overline{Z}$ and standard deviation $s$ of 10 or 20 replicates performed on the same sample on the same day.

Although the measures are made under repeatability condition, this is a misunderstanding of the concept of repeatability as defined in the ISO 5725 standard. When replicates are simply performed on the same day, the total standard deviation may erroneously overestimate or underestimate the method's precision. When the same experiment is repeated, it is frequent to obtain a different value. Therefore, it is compulsory to distribute replicates over several days under intermediate precision conditions, and extract the components of the total variance by applying the ANOVA algorithm to obtain a pertinent estimate of repeatability. Section 3.4 explains how to do this easily with classical worksheet software.

## 3.4 Software Implementation

### 3.4.1 ANOVA Classic Algorithm

Some care must be taken because most statistical software includes an ANOVA procedure. For example, the Excel "Analysis Toolpak" is an add-in containing the option "ANOVA: Single Factor." This can be misleading because there are several ways to perform an ANOVA. In the Excel toolbox, it is assumed that the data is structured with a fixed effect factor.

Whereas for measurement method precision assessment, the model is assumed to be a *random effect factor,* as explained above. The algorithm for fixed effects is different and not adapted to the precision parameter estimation. Fixed effect ANOVA is devoted to making multiple comparisons of means as the factor values are controlled by the experimenter, such as different temperatures. While random factor ANOVA is used to extract the variance components of the dataset. The results obtained with the fixed effect model cannot be used to estimate the between-series or between-laboratory variance, and therefore the precision parameters. But it can be applied to verify if all laboratory means are statistically equal.

If there is a single mean that does not conform with the others, it is said to be an outlier, and the laboratory must be discarded. This is not the procedure implemented in the ISO 5725 standard, and a short presentation of outlier and straggler detection standardized methods is presented at the end of this chapter. Therefore, the Excel statistical add-in cannot be used for computing precision, and it is necessary to develop a specific tool to obtain useful results.

An example using the LEAD dataset is given in the Resource E worksheet. It is designed to manage 11 laboratories or series and 3 replicates/laboratories. It must be adapted to diverse situations where the numbers of participants (or series) and of replicates are larger. Some explanations may help in the implementation of this worksheet. As was the case for Resource A, the formulas are visible in column `C` and column `B` contains the expected result.

Raw data is input in area `B4:E14`, one line par laboratory. The laboratory sum of squares is computed in column `E` with header "$SS_i$." The built-in statistical function `DEVSQ` allows one to directly obtain any sum of squared deviations. Fortunately, if cells are empty, they are considered absent and not put to 0. This means the same worksheet can be used when the number of series or replicates is smaller.

Due to the lack of space, corresponding formulas are not visible, but, for instance, in cell `F4`, this formula was input `=DEVSQ(B4:D4)`, and the result was 0.003800. All $SS_i$ are summed in cell `B19` to give the $SS_W$. This is the direct application of the Eqs. (3.9) and (3.10) for the within-laboratories (or series) sum of squares. The total sum of squares $SS_t$ is obtained in cell `B20` by using the same function but applied to the whole data area `=DEVSQ(B4:D14)`.

In cell `B21`, the calculation of $SS_B$ by difference may pose a problem because the resulting value can be negative, which sometimes happens when all the means are almost equivalent. As explained, the ISO 5725 standard recommends forcing the value of $s_L^2$ (or $s_B^2$) to 0, otherwise, an error occurs when calculating the square root (a negative value having no real square root). The decision rule is simple:

$$\text{If } s_L^2 < 0 \rightarrow s_L^2 = 0$$

It is coded in two steps: first, compute $s_L^2$ in cell `B21` by applying formula (3.13); then check if this value is strictly smaller than 0 with the Excel formula `=IF(B23<0;0;B23)`. As explained later in this chapter, specialized statistical software offers another algorithm called restricted maximum likelihood (REML), which avoids this problem. As this worksheet is devoted to work on a balanced design, it is necessary to verify if this condition is satisfied.

The Excel built-in function `COUNT`, which returns the number of not empty cells, is used to make the control. The total number of non-null data is computed in cell `B17` with `=COUNT(B4:D14)`. The number of laboratories or series is reported in cell `B16` with `=COUNT(B4:B14)` and the number of replicates in cell `F4` with `=COUNT(B4:E4)`. The experimental design must be balanced, and the verification is done in cell `B18` with the formula `=IF(F4*B16<>B17;"Error";F4)`. This is a rather simplistic method of checking, as it may fail if the data layout is not correctly set up, but it is just to show what needs to be done to secure the template.

**Resource E Precision parameters for a balanced design (Excel).**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Resource E: Precision parameters for a balanced design** | | | | | |
| 2 | Lead mg/kg in mussel tissue | | | | | |
| 3 | Laboratories | Replicate. 1 | Replicate. 2 | Replicate. 3 | ni | SSi |
| 4 | Lab 01 | 2.08 | 2.00 | 2.01 | 3 | 0.00380 |
| 5 | Lab 02 | 2.00 | 1.93 | 1.89 | 3 | 0.00620 |
| 6 | Lab 03 | 2.10 | 2.44 | 1.96 | 3 | 0.12187 |
| 7 | Lab 04 | 2.45 | 2.34 | 2.49 | 3 | 0.01207 |
| 8 | Lab 05 | 1.95 | 1.89 | 1.93 | 3 | 0.00187 |
| 9 | Lab 06 | 1.85 | 1.91 | 1.89 | 3 | 0.00187 |
| 10 | Lab 07 | 2.01 | 2.00 | 2.06 | 3 | 0.00207 |
| 11 | Lab 08 | 2.00 | 2.09 | 1.98 | 3 | 0.00687 |
| 12 | Lab 09 | 2.11 | 2.03 | 2.14 | 3 | 0.00647 |
| 13 | Lab 10 | 2.02 | 1.98 | 1.97 | 3 | 0.00140 |
| 14 | Lab 11 | 2.02 | 2.00 | 2.04 | 3 | 0.00080 |
| 15 | **Main parameters** | | | | | |
| 16 | Number of laboratories (I) | 11 | =COUNT(B4:B14) | | | |
| 17 | Number of measurements (IJ) | 33 | =COUNT(B4:D14) | | | |
| 18 | Number of replicates (J) | 3 | =IF(E4*B16<>B17;"Error";E4) | | | |
| 19 | SSW | 0.16527 | =SUM(F4:F14) | | | |
| 20 | SSt | 0.82385 | =DEVSQ(B4:D14) | | | |
| 21 | SSB | 0.65859 | =B20-B19 | | | |
| 22 | Repeatability variance ($s^2r$) | 0.00751 | =B19/(B17-B16) | | | |
| 23 | Temporary between lab variance | 0.01945 | =((B21/(B16-1))-B22)/B18 | | | |
| 24 | Between laboratory variance ($s^2L$) | 0.01945 | =IF(B23<0;0;B23) | | | |
| 25 | Reproducibility variance ($s^2R$) | 0.02696 | =B22+B24 | | | |
| 26 | **Precision** | | | | | |
| 27 | Grand mean | 2.047 | =AVERAGE(B4:D14) | | | |
| 28 | Repeatability std. dev. | 0.0867 | =SQRT(B22) | | | |
| 29 | Between laboratory std. dev. | 0.1395 | =SQRT(B24) | | | |
| 30 | Reproducibility std. dev. | 0.1642 | =SQRT(B25) | | | |
| 31 | Limit of repeatability | 0.2453 | =2.83*B28 | | | |
| 32 | Limit of reproducibility | 0.4647 | =2.83*B30 | | | |
| 33 | Relative Std. Dev. Reproducibility (RSDR) | 8.02% | =B30/B27 | | | |
| 34 | **Grand mean variance ($s^2Z$)** | | | | | |
| 35 | Variance ratio A | 2.59 | =B24/B22 | | | |
| 36 | Coefficient Q | 0.409 | =(B35+1)/(B18*B35+1) | | | |
| 37 | Grand mean variance ($s^2Z$) | 0.0073 | =B25*SQRT(1/(B17*B36)) | | | |
| 38 | Grand mean std. dev. (sZ) | 0.0856 | =SQRT(B37) | | | |

Finally, it is advisable to protect this worksheet against any unintentional changes. It should be remembered that the number of significant figures in a cell depends on its display format and not on the rounding of the data. If data rounding is necessary, use the `ROUND` function but always at the end of the computation. The case of an unbalanced dataset is addressed in the Section 3.4.2, and a modified

worksheet is presented. But it does not offer much difficulty as the application of Eqs. (3.26)–(3.29) requires no other built-in function.

### 3.4.2 Detect Outliers and Stragglers

The initial application of the precision model developed in ISO 5725 standards was for commercial purposes. At present, many regulatory bodies are using it as a prerequisite to accept any analytical method for official control. For example, when a new ISO standard operating procedure is developed, it is mandatory to conduct an interlaboratory study and publish the precision values obtained in an informative annex. As explained, the other application of this standard algorithm is the computation of intermediate precision parameters. It is less promoted, although it is fully presented in standard ISO 5725 and especially useful for Chapters 5 and 6.

The official exploitation of reproducibility has many practical consequences. Unfortunately, exceptionally distant data values may be present and alter the results, and a large part of the ISO standards is devoted to the detection of non-normal values, called outliers and stragglers. This issue is less important when the computation is applied to intermediate precision. When considering Figure 3.2 , three kinds of outliers may be suspected, which are displayed in Figure 3.6. A full collection of statistical tests is proposed to verify several types of rejection hypotheses. They are applied at two values of $\alpha$ error risk:

- If $\alpha > 5\%$ (or $\alpha > 2.5\%$ in some guidelines) the data is a straggler.
- If $1\% > \alpha > 5\%$ (or $0.5\% > \alpha > 2.5\%$) it is an outlier.

The decision rule varies depending on the $\alpha$-value. Given the cost of recruiting a laboratory and the fact that interlaboratory study parameters are only publishable

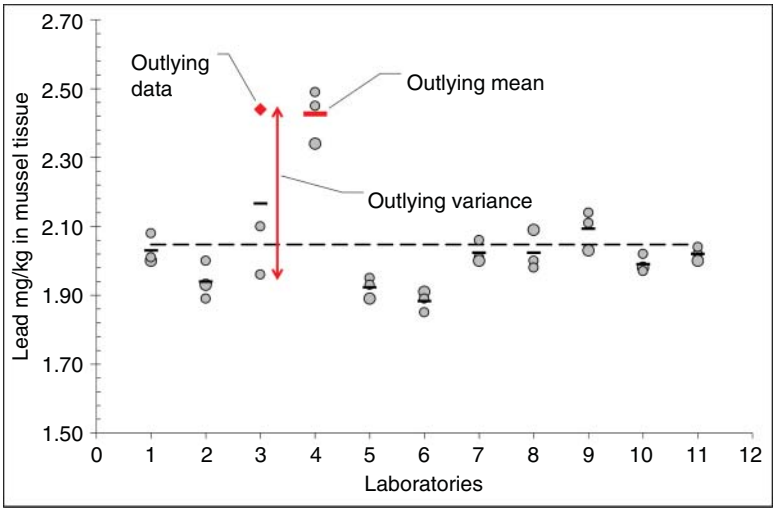



**Figure 3.6** Diverse types of outliers in an interlaboratory study.

when at least eight laboratories obtain acceptable results, it was interesting to consider the possible rejection of one outlying replicate to keep a laboratory. Statistical literature is very abundant on the question of outlier significance testing. Several dozen authors have described tests for rejecting outliers from a dataset.

A limited set of rejection tests was selected by the group of experts in charge of preparing the ISO 5725 standard. Because many outlier rejection tests exist but may not always identify the same data as an outlier, it seemed important to the experts to clearly put a well-identified list in the standard. The complete operating procedure for each particular test is not presented here but is described in the standard for the interested reader. Anyway, the specific task of data polishing is left to the statistician in charge of the study. Apart from Dixon's test, when a laboratory is identified as an outlier, all its data are usually eliminated. The diverse outliers are listed here with their corresponding official tests.

| Hypothesis | Overestimated | Official test |
|---|---|---|
| One outlying variance | Repeatability | Cochran's |
| One or two outlying averages | Reproducibility | Grubbs' single or double |
| One outlying replicate in a lab | Lab variance | Dixon's |

Because outlier rejection is an overly sensitive subject for analysts, many criticisms were issued once the proposed tests were published. Some of these objections are relevant. For example, with Dixon's test, it is possible to declare that one data out of four is an outlier and the distribution is non-Gaussian. This consequence seems strange when it may take more than 1000 measurements to prove that a variable is distributed according to a normal distribution. Depending on the sequence of application of the tests, it is not always the same laboratories that are eliminated, resulting in different estimates of precision.

To reduce this risk of inconsistency, it was also decided to define a standardized testing sequence for eliminating data. The best known is the ISO protocol that defines, on the one hand, a sequence of application of the tests and, on the other hand, an acceptance criterion for the whole study. The criterion consists of calculating the ratio of excluded laboratories to the total number of participants. When more than 2/9 laboratories have been eliminated after the following sequence of tests, the method is not validated, i.e. it is not publishable as a standard or used for commercial purposes. The ISO protocol is as follows:

(a) Cochran's test at 5% risk.
(b) Grubbs' test to cell replicates having a suspect variance; then eliminate outliers.
(c) Go back to (a) if only one replicate is eliminated, otherwise, go to (d).
(d) Cochran's test at 1% risk to all variances and eliminate cells with outliers.
(e) Go back to (a) if only one cell is eliminated, otherwise, go to (f).
(f) Apply simple Grubbs' test at 1% risk to all remaining cell means.
(g) If a mean is eliminated, return to (f), otherwise, go to (h).

(h) Apply double Grubbs test at 1% risk to all remaining cell means.
(i) Calculate the proportion of eliminated laboratories and verify that it <2/9.

The alternative protocol promoted by the International Union of Pure and Applied Chemistry (IUPAC) is similar to the ISO protocol but does not include the Grubbs' test for the elimination of outliers in a laboratory. This sequence of rejection tests is extraordinarily complex and, even when automated, requires manual decision-making. We have presented in some detail this very touchy topic of outliers because it is a repeated concern for analysts. It shows that even a consensus solution among expert statisticians does not result in a very satisfactory answer. The conclusion can be drawn that much attention must be paid to outlier rejection and statistical testing must be applied cautiously.

To illustrate the consequence of the rejection of non-normal data, outliers were eliminated from the LEAD dataset: the complete subset of laboratory 04 and one extreme measure of laboratory 03, as illustrated in Figure 3.7. When doing this operation, the experimental design is no longer balanced. Modification of the initial Resource E worksheet is necessary to manage the new dataset. Modified formulas are presented in Section 3.3 and coded in the new worksheet named Resource F, at the end of the chapter. It is mainly the formula used to compute the between-laboratory variance, which is modified in cell `B25` as reported in Eq. (3.29).

The consequence of the outlier elimination is striking. The reproducibility standard deviation $s_R$ is divided by 30, coming from 0.1642 down to 0.00540 mg/kg, as well as the limit of reproducibility $R$. Now, the largest acceptable difference between the two laboratories is no longer 0.47 mg/kg but only 0.07 mg/kg. It is easy to understand the economic consequences on international trade when the analytical method is used for the official control of Pb in foods. The control laboratories must be more efficient, and consumer health must be better preserved.

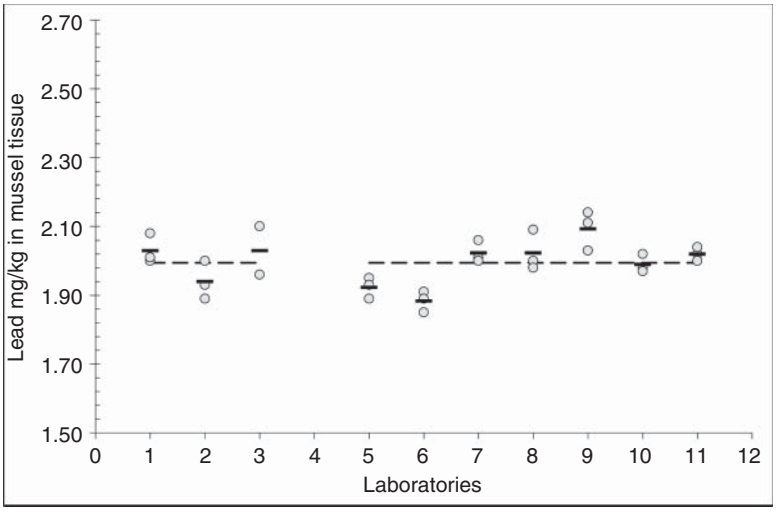



**Figure 3.7** LEAD – interlaboratory study after outlier deletion.

**Resource F Precision parameters for an unbalanced design (Excel).**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Resource F: Precision parameters for an unbalanced design** | | | | | |
| 2 | Lead mg/kg in mussel tissue | | | | | |
| 3 | Laboratories | Rep. 1 | Rep. 2 | Rep. 3 | SSi | $ni^2$ |
| 4 | Lab 01 | 2.08 | 2.00 | 2.01 | 0.003800 | 9 |
| 5 | Lab 02 | 2.00 | 1.93 | 1.89 | 0.006200 | 9 |
| 6 | Lab 03 | 2.10 | | 1.96 | 0.009800 | 4 |
| 7 | Lab 04 | | | | | |
| 8 | Lab 05 | 1.95 | 1.89 | 1.93 | 0.001867 | 9 |
| 9 | Lab 06 | 1.85 | 1.91 | 1.89 | 0.001867 | 9 |
| 10 | Lab 07 | 2.01 | 2.00 | 2.06 | 0.002067 | 9 |
| 11 | Lab 08 | 2.00 | 2.09 | 1.98 | 0.006867 | 9 |
| 12 | Lab 09 | 2.11 | 2.03 | 2.14 | 0.006467 | 9 |
| 13 | Lab 10 | 2.02 | 1.98 | 1.97 | 0.001400 | 9 |
| 14 | Lab 11 | 2.02 | 2.00 | 2.04 | 0.000800 | 9 |
| 15 | **Main parameters** | | | | | |
| 16 | Number of laboratories (I) | 10 | =COUNT(B4:B14) | | | |
| 17 | Number of measurements (IJ) | 29 | =COUNT(B4:D14) | | | |
| 18 | Number of replicates (J) | 2.9 | =B17/B16 | | | |
| 19 | Sum of squared numbers | 85 | =SUM(F4:F14) | | | |
| 20 | Correct measurements number | 26.1 | =B17-(B19/B17) | | | |
| 21 | SSW | 0.04113 | =SUM(E4:E14) | | | |
| 22 | SSt | 0.14492 | =DEVSQ(B4:D14) | | | |
| 23 | SSB | 0.10378 | =B22-B21 | | | |
| 24 | Repeatability variance ($s^2r$) | 0.00216 | =B21/(B17-B16) | | | |
| 25 | Temporary between lab variance | 0.00323 | =(B16-1)*((B23/(B16-1))-B24)/B20 | | | |
| 26 | Between laboratory variance ($s^2L$) | 0.0032337 | =IF(B25<0;0;B25) | | | |
| 27 | Reproducibility variance ($s^2R$) | 0.00540 | =B24+B26 | | | |
| 28 | **Precision** | | | | | |
| 29 | Grand mean | 1.994 | =AVERAGE(B4:D14) | | | |
| 30 | Repeatability std. dev. | 0.0465 | =SQRT(B24) | | | |
| 31 | Interlaboratory std. dev. | 0.0569 | =SQRT(B26) | | | |
| 32 | Reproducibility std. dev. | 0.0735 | =SQRT(B27) | | | |
| 33 | Relative std. dev. Reproducibility (RSDR) | 3.68% | =B32/B29 | | | |
| 34 | **Grand mean variance ($s^2Z$)** | | | | | |
| 35 | Variance ratio A | 1.49 | =B26/B24 | | | |
| 36 | Coefficient Q | 0.468 | =(B35+1)/(B18*B35+1) | | | |
| 37 | Grand mean variance ($s^2Z$) | 0.0015 | =B27*SQRT(1/(B17*B36)) | | | |
| 38 | Grand mean std. dev. (sZ) | 0.0383 | =SQRT(B37) | | | |

#### 3.4.2.1 Other Algorithms

As shortly explained, it is common to encounter several problems when eliminating outliers. For example, if an obvious outlier (for the analyst) is not detected, too many laboratories are excluded, the study cannot be published, the precision is underestimated, etc. Finally, the estimates of repeatability and reproducibility are not satisfactory. For all these reasons, the ISO 5725 standard includes a set of *robust estimators* that have the advantage of being weakly affected by the presence of outlying data. The median is the typical example of a robust parameter for statisticians. In this context, the meaning of the word robustness is slightly different from that used in evaluating the analytical method's robustness. Finally, when robust estimators are used, it is pointless to eliminate suspect data.

The principle of the robust algorithms proposed by ISO is to iteratively transform the data by using robust location and dispersion estimators. Thus, it is possible

to calculate robust versions of the precision parameters, namely $\overline{\overline{Z}}$, $s_r$, and $s_R$. The application of these algorithms to interlaboratory studies is not developed here. It is mainly of interest to professional interlaboratory study organizers. For the details, it is possible to refer to Part 5 of the ISO 5725 standard or to the brief published by the Analytical Methods Committee of the Royal Society of Chemistry [6, 12, 13].

However, an example of a robust parameter applied to proficiency testing schemes is described in Section 4.3.4 and Resource G; it is officially recognized as Algorithm A. It gives an insight into the principle of this more recent statistical method. This evolution is pointed out here because it marked the introduction of new statistical methods based on iterative calculation in standards. Experience shows that it is interesting to compare the precision values obtained by the classic calculation method, before and after eliminating outliers, with those of the robust algorithm. In general, obtained values are in-between, which probably better reflects reality. However, the standardized robust methods are not fully satisfactory because they opaquely replace *problematic* data with other values closer to the rest of the dataset in such a manner that data traceability is lost.

## References


1 Bureau International des Poids et Mesures (BIPM) (2012). *International Vocabulary of Metrology — Basic and General Concepts and Associated Terms (VIM3)*, JCGM 200:2012. BIPM https://www.bipm.org/ (accessed 23 July 2023).

2 Standard ISO 5725-1:1994 *Accuracy (Trueness and Precision) of Measurement Methods and Results — Part 1: General Principles and Definitions*. Genève: ISO.

3 Standard ISO 5725-2:2019 *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*. Genève: ISO.

4 Standard ISO 5725-3:1994 *Accuracy (Trueness and Precision) of Measurement Methods and Results — Part 3: Intermediate Measures of the Precision of a Standard Measurement Method*. Genève: ISO.

5 Standard ISO 5725-4:2020 *Accuracy (Trueness and Precision) of Measurement Methods and Results — Part 4: Basic Methods for the Determination of the Trueness of a Standard Measurement Method*. Genève: ISO.

6 Standard ISO 5725-5:1998 *Accuracy (Trueness and Precision) of Measurement Methods and Results — Part 5: Alternative Methods for the Determination of the Precision of a Standard Measurement Method*. Genève: ISO.

7 Feinberg, M., Montamat, M., Rivier, C. et al. (2002). Comparison of strategies to quantify uncertainty of lead measurements in biological tissue at mg $kg^{-1}$ level. *Accreditation and Quality Assurance* 7: 403–408.

8 Sahai, H. and Ojeda, M.M. (2004). *Analysis of Variance for Random Models, Volume 1: Theory, Methods, Applications, and Data Analysis*. Boston, MA: Birkhäuser.

**9** Sahai, H. and Ojeda, M.M. (2005). *Analysis of Variance for Random Models, Volume 2: Unbalanced Data: Theory, Methods, Applications, and Data Analysis*. Springer-Verlag.

**10** Bugner, E. and Feinberg, M. (1992). Determination of mono-and disaccharides in foods by interlaboratory study: quantitation of bias components for liquid chromatography. *Journal Of AOAC International* 75 (3): 443–464.

**11** Rozet, E., Ceccato, A., Hubert, C. et al. (2007). Analysis of recent pharmaceutical regulatory documents on analytical method validation. *Journal of Chromatography A* 1158: 111–125.

**12** Analytical Methods Committee of the Royal Society of Chemistry (AMC) (1989). Robust statistics: how not to reject outliers: part 1. Basic concepts. *Analyst* 114: 1693–1697.

**13** Analytical Methods Committee of the Royal Society of Chemistry (AMC) (1989). Robust statistics: how not to reject outliers: part 2. Interlaboratory trials. *Analyst* 114: 1699–1702.

# 4

# Trueness

## 4.1 Trueness and True Value

The Bureau International des Poids et Mesures (BIPM) oversees developing and promoting metrology, which is "the set of techniques and know-how that allow measurements to be made and to have sufficient confidence in their results." It is therefore one of the keys to assessing trueness. Analytical chemists often stay away from metrology, which represents a considerable challenge in terms of science, economics, and societal demands. This can be explained by the fact that analyst concerns were only recently addressed by metrologists. Several tools developed by the BIPM to assess trueness in analytical sciences are presented in Section 4.2 but for many decades, the question of measuring the "amount-of-matter" was neglected by the BIPM. Recently, a strong effort has been made, and many solutions are now proposed.

While there is a continuum of parameters for expressing precision, ranging from repeatability to reproducibility through intermediate precision standard deviations, the estimation of trueness is much simpler, in theory. Whatever the parameter used, they all represent the same idea:

> "the closeness of agreement between the average of an infinite number of replicate measures of a quantity values and a reference value."

Although the International Vocabulary of Metrology (VIM) is very explicit, the term trueness is sometimes confounded with accuracy. Accuracy is the "closeness of agreement between a measured value and a true value of a measurand." According to the VIM philosophy, accuracy is a concept and is not quantifiable, while trueness is more practical and measurable. This difference is better explained in Section 5.1.1 about the vocabulary used in method validation.

The challenge in assessing trueness is to obtain the adequate reference value or true value, which is denoted $X$ throughout this book. For some analysts, this is such a tricky issue that they may think it is impossible, for instance, when the analyte is defined by the method itself, as it is in food microbiology.

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

### 4.1.1 Bias and Recovery Yield

Trueness raises the question of the definition of the analyte. It was partly discussed in Chapter 1 about quantification. To introduce trueness parameters, let us keep the notations already used, namely:

- $X$ the target (or reference or assigned) value of a sample.
- $\overline{Z}$ the average of replicates $Z_i$ obtained for this sample.

The three parameters listed here are common to express trueness, They do not really measure the trueness but rather the lack trueness. This is why they are called as bias, except the recovery yield

Bias

$$\delta = \overline{Z} - X \tag{4.1}$$

Relative bias

$$\delta\% = \left(\frac{\overline{Z} - X}{X}\right) \times 100 \tag{4.2}$$

Recovery yield

$$RY\% = \frac{\overline{Z}}{X} \times 100 = 100 - \delta\% \tag{4.3}$$

They are roughly equivalent. If the value of the bias or relative bias is negative, it means that the method underestimates the target concentration, but if it is positive, the method overestimates. For the recovery rate, a comparison must be made with respect to 100%. For analysts who are pleased to utilize percentages, the recovery rate is the most meaningful. The relative bias is the complement of the recovery yield and can be interpreted as a correction factor that measures the proportion of what was measured compared to what was expected to be measured.

Most regulatory documents also propose percentages as acceptance criteria. Similarly, precision can be expressed as a percentage by estimating relative standard deviations, as explained in Section 3.2.1; they define trueness by means of the recovery yield. An example is given in Section 7.5.1.

Below, the worksheet excerpt illustrates these parameters and data in cells `B5` and `B6`, showing that in the worksheet it is not useful to multiply results by 100 to obtain a percentage. The editing format "Percent" directly gives the expected value and adds the "%" symbol. This mode of expression is preferable because the parameter may be used later, for instance, to correct a measurement value. It is easier as it is not necessary to multiply or divide the data by 100.

| | A | B | C |
|---|---|---|---|
| 1 | **Trueness** | | |
| 2 | Reference value | 14 | |
| 3 | Measurement value | 13.6 | |
| 4 | Absolute bias | –0.4 | =B3 - B2 |
| 5 | Relative bias | –2.86% | =(B3 - B2)/B2 |
| 6 | Recovery yield | 97.14% | =B3/B2 |

For some methods it is conventional to transform measurements into decimal logarithms, for instance, the microbiological counting methods. In this case, absolute

bias is equivalent to the recovery yield since the logarithm of a ratio is equal to the difference of the logarithms.
Recovery yield in log

$$\log\left(\frac{\overline{Z}}{X}\right) = \log(\overline{Z}) - \log(X) \tag{4.4}$$

### 4.1.2 Evolution of the Concept of True Value

In the 1980s, BIPM decided to address the question of measurement error and proceeded from what is called the traditional "error approach" to the new uncertainty approach, and to reconsider some extensively used concepts.

The rationale behind the error approach was to determine an estimate of the true value that was as close as possible to the theoretical true value. This was based on the previous definition of accuracy. In this context, the true value was considered unique and, in practice, unknowable. It was assumed that the deviation from the true value was a combination of different sources of error. To differentiate the sources of error, the concepts of fixed error and random systematic error were introduced as always distinguishable and to be treated differently. For example, systematic error may be due to the use of poorly graduated glassware, while random error is caused by a drifting measuring device that does not always give the same value.

In fact, it is often recognized a systematic error can be claimed as fixed when its variations in amplitude are small when compared to the measurement value, but it becomes random when they are large enough to introduce measurable fluctuations. It is traditional to say that the categories of systematic error vary according to the principles of the method. Controlling them is part of the analyst's expertise: the development of the method must make it possible to highlight and correct them. Thus, a bias can have various origins, such as:

- Evolution of the sample over time.
- Degradation of calibration solutions.
- Modifications of equipment settings.
- Miscalculations.
- Human errors.

Figure 4.1 illustrates a source of systematic error frequently encountered in chromatography. It could be related to the integration method used to calculate the areas of poorly resolved peaks. In the graph, two peaks are represented, centered on retention times of 2.0 and 3.1 min, respectively, with a drift of the baseline. The peak of interest is the second, smaller peak.

The method used is the drop-line algorithm, symbolized by the dashed lines. It consists in locating the beginning and end of each peak, considering the drift, and then assigning to each one a part of the common triangular overlapping area, by drawing the perpendicular to the point of separation of the peaks. In the graph on the left, since the surfaces of the two peaks are quite close to each other, the error remains negligible. However, it becomes significant when the area of the peak of interest decreases, as in the graph on the right.

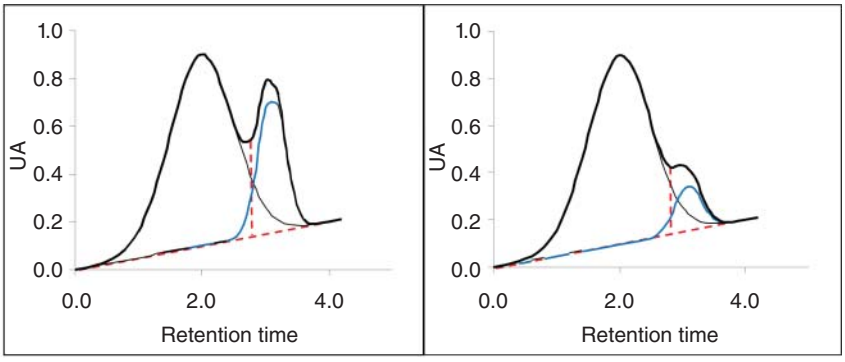



**Figure 4.1** Example of systematic error generated by the integration mode of poorly resolved chromatographic peaks.

Furthermore, if the major peak is very asymmetric, the situation becomes worse. The uncertainty approach is to recognize that, owing to the inherently incomplete amount of detail in the definition of a quantity:

> "there is not a single true value but rather a set of true values consistent with the definition."

However, the complete set of values is, in principle and in practice, unknowable. However, it is possible to estimate an interval containing a given proportion of these true values. The change in the paradigm of the single, unique true value to the bounds of an interval is an especially important fact for experimenters. More details on the consequences are described in Sections 4.3.2 and 6.3.

### 4.1.3 Specificity and Sources of Bias

In the context of analytical sciences, specificity is "the capability of a measuring system or operating procedure to measure the concentration of a given analyte." The main problem when addressing specificity is the potential presence of interfering compounds that alter the analyte's detection integrity. Analytical literature is extensively dedicated to the topic of the various chemical, physical, or biological mechanisms that are the source of interference and specific to each method. For example, in atomic absorption spectrophotometry, spectral interferences can be due to the lack of resolution of monochromators, or the broadening of the absorption bands caused by the earth's magnetic field and matrix interferences to various elements absorbing at the same wavelength as the searched analyte.

Figure 4.2 is a classic representation of bias, which is made possible when considering how the inverse-predicted concentration is computed (Section 2.2). First, let us define the first bisecting line (dotted line) as the *trueness line* where measurements are not biased, i.e. the inverse-predicted concentration is exactly equal to the known reference concentration. Then consider a collection of materials whose reference values are known. Once they are

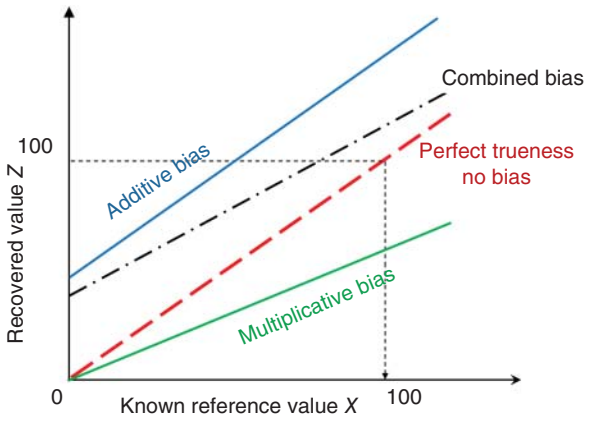



**Figure 4.2** Geometric interpretation of additive and multiplicative bias.

analyzed, the inverse-predicted concentrations are reported on the graph for each material. For instance, such a diagram can be obtained with a single material receiving different spikes as explained in Chapter 1 about the standard addition method (SAM). Figure 4.2 illustrates three hypothetical situations:

- Measurement values are systematically upwards shifted in concentration, and the recovery line is shifted by an additive bias (it can be negative, but it still is an additive bias).
- Measurement values are proportionally diverging from the concentration, and the slope of recovery line is modified by a multiplicative bias.
- Both additive and multiplicative biases are combined.

Keeping the same notation, i.e. $X$ the reference value and $Z$ the inverse predicted concentration, these three hypothetical situations can be modeled as follows:

| | |
|---|---|
| Trueness line (no interferences) | $Z = X$ |
| Multiplicative bias | $Z = b_1 X$ |
| Additive bias | $Z = b_0 + X$ |
| Combined bias | $Z = b_0 + b_1 X$ |

Unfortunately, it is impossible to experimentally isolate each type of bias, as they are usually combined. But if we assume that the model connecting $X$ and $Z$ is a straight line, the recovery line can be used.

- To check if the trueness is linear or varies in a more complex way with the concentration.
- To estimate the coefficients of the model $Z = b_0 + b_1 X$ and define an eventual correction factor. Examples of correction factors are presented in Sections 8.4.2 and 10.2, with consequences for measurement uncertainty (MU).

When it is not possible to globally evaluate the matrix interferences, the SAM represents a convenient technique to highlight and monitor the interferences; possible procedures are presented in Chapter 1.

Because trueness is defined by reference to a true value, it will be assumed in the rest of this chapter that there is one true value.

## 4.2 Assessment of Trueness

### 4.2.1 Primary Operating Procedures

To bring some enlightenment, it is interesting to consider how official metrology bodies and BIPM establish reference values. The BIPM covers nine metrological areas, among them chemistry and biology, i.e. the analytical sciences in general. The other areas are acoustics, ultrasound and vibration, electricity and magnetism, length, mass and related quantities, photometry and radiometry, ionizing radiation, time and frequency, and thermometry. There are related specialized consultative committees that elaborate the way in which the units and the related standards are defined. For the analytical sciences, this is the Consultative Committee for Amount of Substance, or Consultative Committee for the Quantity of Matter (CCQM), one of the latest committees to be created since it was only established in 1993 while other committees already existed.

The difficulty of defining international standards in this area can explain its postponed creation. The task was easier for other committees, while concrete standards, such as the meter or the kilogram, have existed for a long time. To overcome these shortcomings, during the first CCQM meeting in 1993, when chemistry was recognized as a separate field of metrology, it was decided to introduce the primary methods of measurement, defined as a "reference measurement procedure used to obtain a measurement result without relation to a measurement standard for a quantity of the same kind."

In other words, a primary method of measurement allows a quantity to be measured in terms of a particular International System SI unit without reference to a standard or measure already expressed in that unit. In principle, it is completely independent of the measurements of the same quantity but calls upon measurements expressed in other units of the SI.

Unfortunately, only a few methods of analysis meet that definition. It is mainly coulometry, gravimetry, titrimetric, freezing point depression, and isotope dilution with mass spectrometry (ID-MS). Considering that chromatographic equipment represents more than 80% of the world's instrumental fleet, there is still a long way to go before routine methods are linked to primary methods. As examples of primary standards, VIM cites:

- A primary standard prepared by dissolving a known quantity of a chemical substance in a known volume of solution; typically, a calibration solution.
- A standard for isotope molar ratio measurements, prepared by mixing known amounts of specified isotopes, as a reference material (RM) used for isotopic dilution.

Recently, numerous studies have been launched to validate other methods as primary.

### 4.2.2 Reference Materials

The difficulty of tracing chemical measurements to SI units, linked to the absence of a unique international material standard for the mole, has been highlighted by CCQM. To overcome this problem, it was decided to develop different RMs. The number of RMs is theoretically infinite, because they are supposed to be available for any molecule in any matrix at various concentrations!

In recent decades, however, there has been a huge increase in their number and interest. Therefore, the increasing number of RMs in analytical chemistry would make an exhaustive list of them obsolete at the time of writing. Nevertheless, several types can be distinguished, depending on their intended use. The strengths and weaknesses of each are summarized in the following list.

#### 4.2.2.1 Certified Reference Materials (CRM)

The reference value is established by a specialized collaborative study involving metrology reference laboratories. The sources of error are carefully controlled and reduced, but despite this, the MU can still be remarkably high compared to the specifications sought. As they are limited in number, their most typical disadvantages are that the matrix used may differ from the one in the scope of the method; they are not available at the desired concentration level; and their price is high and prohibits any daily use.

#### 4.2.2.2 External Reference Materials (ERM)

They have recently appeared on the market. The multiplication of proficiency tests makes it possible to propose the use of surplus samples for trueness verification or calibration purposes. Their reference value and the corresponding uncertainty are obtained at the end of the test, and their traceability can, depending on the case, be traced back to certified reference materials (CRMs). They have the advantage of being more readily available, less expensive than CRMs, and better adapted to the various needs of laboratories.

#### 4.2.2.3 Internal Reference Materials (IRM)

The user determines the reference value. This may be done in collaboration with other laboratories or by combining methods. The matrix used is well suited to the laboratory's field of application, and it is generally cheaper. The sources of error can still be important and their traceability to SI can be insufficient. On the other hand, they are perfectly adapted to control charts.

#### 4.2.2.4 Verification Standard Solutions

They are prepared independently of the calibration solutions. They do not allow the detection of matrix effects. They are independent of the analytical method and inexpensive. They can also be certified either by a metrology organization or by the reagent supplier.

#### 4.2.2.5 Standard Addition Method (SAM) and Surrogate Samples

Both of these methods are described in more detail in Chapter 1 and come with the different available protocols. Surrogate samples can only be considered correct substitutes for internal validation when the nature of the matrix is perfectly known, as in the case of excipients for drugs. They are cheap, but their traceability is often poor.

ISO 17025 requires for an accredited laboratory that, wherever possible, the traceability of RMs to SI units or CRMs be established. Internal reference materials (IRMs) shall be verified to the extent technically and economically feasible. Although CRMs are the most elaborate form of RM, there are several problems associated with their use.

A few years ago, RMs were a rare commodity. Now, many organizations produce or distribute RMs. Around 1995, the Code of Reference Materials or COMAR database was developed by the French National Bureau of Metrology (BNM); today the German BAM, or Bundesanstalt für Materialforschung und -prüfung is managing https://www.comar.bam.de. It contains tens of thousands of RM references, manufactured in 27 countries by over 200 producers. The European research program called Virtual Institute for Reference Materials (VIRM) also developed a database of approximately 20,000 CRMs. For a long time, the certification procedure for a CRM was conducted in a non-transparent manner.

The allegation "certified" was therefore used somewhat abusively and had not much to do with the third-party *certification* process defined in the ISO 9000 standards. However, there is no single approach to setting a certified reference value. Often, one or more collaborative studies are performed, but by crossing analytical techniques and selecting a few expert laboratories, to identify major biases. However, some organizations believe that it is preferable to entrust this task to a single highly specialized laboratory that will perform the certification alone.

Of course, the economic aspect is mostly considered when making the choice. Sometimes the recommendations of the ISO 5725 or ISO 13528 standards are far from being followed, even if the statistical exploitation is in agreement with them. Great attention must be paid to the preparation of a CRM, the control of its homogeneity (Section 4.3.5) and its shelf-life. In the case of perishable products, dried or freeze-dried materials are used. For powdered materials, the packaging is made with grain-to-grain filling systems. Sometimes modified-atmosphere packaging is used to stabilize the analytes. An RM must have the following qualities:

- Stability over time.
- Homogeneity of the analyte to be determined.
- Concentration in accordance with the usual applications.
- Availability in sufficient quantity.
- Contain several analytes of interest.

These qualities are particularly critical in the case of biological materials, such as foods or biological fluids. Biological matrices are essentially perishable, and their shelf-life is strictly related to their water content. Unstable analytes such as vitamin D in a plant, since it is photosensitive, has not the same shelf life as traces of cadmium in an alloy. CRMs are intended for use by a single laboratory to conduct an in-house

validation study to verify that their results agree with the certified reference value. Unfortunately, compared to the wide variety of matrices processed by analysts, only a limited number of CRMs are available.

Paradoxically, there is no real recommendation on how to compare the result obtained by the laboratory on a CRM with its certified value. As all these materials are now distributed with the MU on their reference value, a simple method could consist in checking if the result obtained is contained in the coverage interval, including the laboratory's own MU. Section 8.2 on sample conformity assessment gives a detailed explanation of such a comparison procedure.

A preliminary precaution is to check the mode of expression of CRM uncertainty, as standard uncertainty $u$ or an expanded uncertainty $U$, knowing that, unless otherwise indicated, $U = 2u$. If $X$ is the reference value, $U(X)$ its expanded uncertainty, $Z$ the result obtained by the laboratory, and $U(Z)$ its uncertainty, the following test is a simple decision rule.

Decision rule for CRMs

$$\frac{|Z - X|}{\sqrt{U^2(Z) + U^2(X)}} \leq 1.96 \tag{4.5}$$

When possible, a more efficient method consists of including CRMs as validation materials for building a method accuracy profile (MAP) and directly checking if the bias and tolerance intervals lie in the acceptance interval (Section 5.1). For accreditation bodies, the value of CRMs is unquestionable, so their use is widely recommended. However, in addition to the difficulties inherent in their conservation, many more fundamental analytical problems persist. For example, there are no CRMs for some official methods, such as the Kjeldahl nitrogen method, because its reproducibility is unsuitable for certification. In addition, very often for such direct methods where the analyte is defined by the method, there is no exact molecular equivalence between what is sought and what is measured.

## 4.3 Proficiency Testing

Fortunately, for those fields of application that do not have CRMs, or do not yet benefit from the work of official metrology organizations nor primary methods, proficiency testing is a possible method for assessing trueness. The classic experimental approach consists of interlaboratory comparison defined as the "organization, performance and evaluation of measurements or tests on the same or similar items by two or more laboratories in accordance with predetermined conditions." Traceability is achieved to a lesser extent than with CRMS but is quite satisfactory, given the shortcomings. The organization and design of collaborative studies vary according to the objectives. These differences do not always appear clearly in the literature. Two major types of collaborative studies can be distinguished.

### 4.3.1 Interlaboratory Comparison or Proficiency Testing Scheme (PTS)

It involves many participants (up to several thousand) who do not repeat their measurements. The objective is to rank the laboratories using different scores and verify

their competence for a given analyte or method, in terms of trueness. They are routinely used by accredited laboratories, which are summoned to participate at several proficiency testing scheme (PTS) per year to remain accredited. More details are available in this chapter.

It is quite different of the classic interlaboratory ring study described in Section 3.1 about precision. It brings together only some laboratories which must perform replicates. It is mainly devoted to estimating the repeatability and reproducibility of a method, but ISO 5725 standards also describe several extensions and parameters relating to trueness. The major disadvantage is the cost in such a way that interlaboratory studies are not routinely used for this specific purpose. This topic is not addressed here.

### 4.3.2 Organization of Proficiency Testing Schemes

While ISO 5725 interlaboratory analyses are organized at the initiative of a limited group of laboratories that want to standardize an operating procedure according to standard recommendations, PTSs are managed by specialized organizations that follow the rules available in ISO standards and guides:

| Reference | Topics covered |
|---|---|
| ISO/IEC Guide 43-1 | Implementation and organization of PTS |
| ISO/IEC Guide 43-2 | Relationship between PTS and certification bodies |
| ISO 13528 | Statistical methods for the interpretation of PTS |
| ISO 17043 | Requirements for PTS organizers |

Usually, the legal status of the managing organization is nonprofit and issued by a professional association, such as cereal products or clinical biology. But it can also be a governmental body if regulatory purposes are involved. In some cases, participation in a PTS is a regulatory requirement. For example, biomedical analysis laboratories must participate in quality control (QC) programs to keep their right to make analyses on humans. To the extent that organizers can obtain certification, it is essential that they have no conflict of interest with participants.

There are many PTS organizers around the world. At the European level, the European Proficiency Testing Schemes (EPTIS) is managing a database available on the site http://www.eptis.bam.de that lists several hundred PTSs in all fields of analysis. Similarly, the US National Association for Proficiency Testing (https://proficiency.org/) and Proficiency Testing Canada (https://ptcanada.org/) play the role of PTS providers in their own countries. The participants can be several hundred or thousands, usually contributors to the nonprofit managing organization. In that context they participate in PTS items selection and interpretation procedures. Each year they receive one or several test samples that they measure with their own routine method. One single measurement value is required but it is not a general rule.

One PTS often covers several analytes. The results are sent back to the organizer, who will calculate a reference value with acceptance limits. The participating laboratory must be within these limits to be declared competent, as required by several accreditation bodies. Finally, a PTS can be summarized in four steps:

1. Select and prepare a homogenized test material packaged in adapted containers. It is usual that they only have the quantity sufficient for the expected number of measurements.
2. Send one item to each member of a panel of subscribing laboratories.
3. Estimate the assigned reference value with its acceptance limits according to one of the consensus calculation methods described in the next section.
4. Compute a score for each participating laboratory and decide whether it is acceptable; the agreement depends on the value of the score.

The principle is therefore quite simple but the organizational and logistical aspects are crucial, justifying the implementation of a quality system for PTS organizers. Especially since they are service providers for accredited laboratories. The organizer's competence mainly lies in the ability to produce homogeneous test materials. Although this is an obvious technical problem, it is impossible to give general rules here, given the variety of matrix types available and the types of analytes. Some materials combined with certain analytes can be very unstable, such as microorganisms in water or foods, while others, such as lead isotopes in rocks, are very stable.

Therefore, any conceivable preservation technology can be used: drying, freeze-drying, canning, pasteurization, or other as well as all types of packaging: inactinic, vacuum, controlled atmosphere, and so on. Most materials are shipped in such quantities so as to avoid the laboratory making additional replicates and no longer remaining in routine conditions. Often, once the package is opened, the contents can change, and the material should not be reused. In addition to the planning aspects, the interpretation of proficiency tests raises statistical problems at three levels:

1. Select an algorithm to estimate the reference value that will be assigned to the material at the end of the PTS. This question of defining the central value of sampled data is as old as statistical science itself. Since this material can eventually be used as an external RM, it is also important to know how to express the MU.
2. Check material homogeneity and stability. If the material is not homogeneous, there is a risk that a laboratory's result will be interpreted as being due to its incompetence. The recommended method is to achieve homogeneity checking at the beginning, but it can be complemented by using the results the laboratories sent back.
3. Define a score of competence for a laboratory. The aim is to establish a certificate that can be provided to an inspector in charge of auditing the laboratory.

### 4.3.3 Reference Value of the Test Material

Some years ago, the influence of various estimators used by different organizers on laboratory scoring was demonstrated [1]. The situation has changed since

**Table 4.1** Terms used in ISO 13528.

| | | |
|---|---|---|
| Assigned value | $X$ | Value attributed to a particular property of a proficiency test item. It is also referred to as the reference value. |
| Standard deviation for proficiency assessment | $s_A$ | The measure of dispersion used in the evaluation of the results of proficiency testing based on the available information. It is comparable to a standard deviation interpreted as a target dispersion value for a population of laboratories. |

already-quoted international standards were established. Statistical data processing is no longer left to the free choice of the test organizers, although the proposed estimators are flexible. Two new concepts were introduced in the standards, as defined in Table 4.1.

Whereas ISO 13528 is applied in many fields of measuring sciences, the definitions are confusing as they try to merge different methodological frameworks. Five estimators can be used for assigning the reference value $X$ to the test material. Each method comes with an estimate of the standard-uncertainty $u(X)$.

The Table 4.2 summarizes these methods. It is assumed that $I$ laboratories are participating in the PTS with $1 < i < I$.

The robust estimates obtained by algorithm A are described in Section 4.3.4. The reason for robust estimators in PTS to be proposed can be explained as no laboratory can be rejected as an outlier, since it is compulsory to give a score to every participant,

**Table 4.2** Estimators of the assigned value and their standard-uncertainty.

| **Method** | **Nature of the test material** | **Standard-uncertainty** |
|---|---|---|
| Formulation | Surrogate matrix or standard addition | Law of propagation of uncertainty (Section 6.6) |
| Certified value | Certified reference material | $u(X) = u(X_{CRM})$ |
| Calibration by CRM | Require simultaneous measures of $X_i$ and $C_i$ on the test material and CRM, respectively. Calculate the average difference. $\overline{D} = \frac{\sum_i (X_i - C_i)}{I}$ | |
| | $X = X_{CRM} + \overline{D}$ | $u(X) = \sqrt{u^2(X_{CRM}) + u^2(\overline{D})}$ |
| Expert laboratories | A group of $I'$ expert laboratories provide $X$ and their uncertainty $u(X_i)$ with $1 < i < I'$ | $u(X) = \frac{1.23}{I' \sqrt{\sum_i u^2(X_i)}}$ |
| All laboratories | Algorithm A: $X$ robust average and $s_A^*$ robust standard deviation (Section 4.3.5) | $u(X) = \frac{1.23 \times s_A^*}{\sqrt{I}}$ |

**Table 4.3** Estimators of the standard deviation for proficiency assessment.

| Method | Principle | Formula |
|---|---|---|
| Prescribed value | Regulatory requirement | Not applicable |
| From a model | e.g. the Horwitz's model | $s_A = 0.02X^{-0.849}$ |
| By perception | Prerequisite performance level | $s_A = \frac{X}{3}$ |
| Interlaboratory study | Based on $s_r$, $s_R$ | $s_A = \sqrt{2.25\left(s_R^2 - s_r^2\right) + \frac{s_r^2}{I}}$ |
| Proficiency testing | Algorithm A (Section 4.3.5) | $s_A^*$ |

even when it is an obvious outlier. Moreover, statistical testing for outlier rejection always assumes a simple distribution function behind the dataset, usually Normal or unimodal. Experience shows that the situation is more complex as bimodal or more distributions may occur.

For the standard deviation $s_A$, five methods are also available and reviewed in Table 4.3, including the robust calculation with algorithm A described below. The organizer must choose among these methods, in consultation with the members of the PTS, an eventual accreditation body, or the regulatory requirements.

### 4.3.4 Performance Scores

A performance score is applied to measure the extent to which a given laboratory diverges from the others and to judge whether it has the same competence as the rest of the group of laboratories that participated in the test. Therefore, a score must be accompanied by an acceptance interval to allow this judgment. Table 4.4 combines the five calculation methods proposed by the ISO standards, with their acceptance limits. The first two scores are just usual measures of bias. The $z$-score is the best-known and most widely used. It consists of centering the result obtained (by

**Table 4.4** The different scores available for PTS.

| Designation | Formula | Acceptance |
|---|---|---|
| Absolute bias | $D_i = Z_i - X$ | $\pm 3\, s_A$ |
| Relative bias | $D\%_i = \frac{Z_i - X}{X} \times 100$ | $\pm 3\, \frac{X}{s_A}$ |
| $z$-score | $z_i = \frac{Z_i - X}{s_A}$ | $\pm 3.0$ |
| Standardized deviation $EN$ | $EN_i = \frac{Z_i - X}{\sqrt{U^2(Z_i) + U^2(X)}}$ | $\pm 1.0$ |
| $z'$-score | $z'_i = \frac{Z_i - X}{\sqrt{s_A^2 + u^2(X)}}$ | $\pm 3.0$ |

subtracting the mean) and reducing it (by dividing it by the standard deviation) to find a centered and reduced Normal distribution. The advantage of this operation called *standardization* is to avoid scale effects that exist with absolute bias and allows comparison of the laboratory's performance for various tests, regardless of the units used or the concentration level.

Understanding the principles used to determine the acceptance interval does not require many comments. For example, for the $z$-score, it is referring to the standard Normal distribution law, and $\pm 3$ limits correspond to a probability interval of 99.7%. Proposed score must be interpreted separately, in the framework of a given PTS. When a laboratory participates in several PTS, it would be an error to compute some average value because each score is linked to a specific set of values.

### 4.3.5 Algorithm A

The existence of robust statistics, as defined by statisticians, was already introduced in Section 3.4.2 about the possible presence of outliers. The theory of this new part of statistics was developed in the 1980s, and many robust parameters have been published since. They are good reference books, but this is not the right place to review possible proposals [2]. The aim is to only give an insight into a robust estimation method, introduced by ISO, in relation to the assessment of the PTS performance. The results of a proficiency test that involved determining moisture in alfalfa are used to illustrate the interest in so-called algorithm A, which simultaneously gives robust estimates of mean and standard deviation. Thirty-six laboratories, labeled L01 to L36, participated and their results are listed in Table 4.5.

Algorithm A is a full part the of ISO 13528 standard. Resource G explains how the algorithm works and how robust parameters are iteratively computed until a convergence criterion is reached. Because it is an iterative algorithm, it is simpler to use a Python script but it is programmable in Excel with Visual Basic for Applications.

**Table 4.5** ALFALFA[a] – proficiency testing of moisture determination (g/100 g).

| Lab | $z_i$ | Lab | $z_i$ | Lab | $z_i$ | Lab | $z_i$ |
|---|---|---|---|---|---|---|---|
| L01 | 7.59 | L10 | 8.06 | L19 | 8.17 | L28 | 8.27 |
| L02 | 7.79 | L11 | 8.07 | L20 | 8.18 | L29 | 8.28 |
| L03 | 7.81 | L12 | 8.07 | L21 | 8.19 | L30 | 8.29 |
| L04 | 7.81 | L13 | 8.10 | L22 | 8.20 | L31 | 8.29 |
| L05 | 7.84 | L14 | 8.11 | L23 | 8.21 | L32 | 8.30 |
| L06 | 7.86 | L15 | 8.15 | L24 | 8.22 | L33 | 8.31 |
| L07 | 7.92 | L16 | 8.15 | L25 | 8.23 | L34 | 8.31 |
| L08 | 7.96 | L17 | 8.15 | L26 | 8.27 | L35 | 8.36 |
| L09 | 7.99 | L18 | 8.16 | L27 | 8.27 | L36 | 8.57 |

a) Unpublished personal data.

**Resource G Algorithm A (Python).**

Import package `numpy` (for numerical Python) which contains useful mathematical functions.

```
import numpy as np
```

Moisture in alfalfa data is stored in a numerical array. This conversion is useful for applying `numpy` functions.

```
Z = np.array([7.59, 7.79, 7.81, 7.81, 7.84, 7.86, 7.92,
7.96, 7.99, 8.06, 8.07, 8.07, 8.1, 8.11, 8.15, 8.15,
8.15, 8.16, 8.17, 8.18, 8.19, 8.2, 8.21, 8.22, 8.23,
8.27, 8.27, 8.27, 8.28, 8.29, 8.29, 8.3, 8.31, 8.31,
8.36, 8.57])
```

Compute the classic parameters for comparison with robust.

```
print("Classic")
print(np.mean(Z), np.std(Z))
```

The median of the observed values $Z_i$ is used to obtain the value `max_dist` that is used to detect outliers. It is the median of the absolute deviates to the median multiplied by the correction factor 1.483. This intensive use of the median is frequent for robust estimators. The correction factors were determined by simulation [3, 4].

```
Z_median = np.median(Z)
max_dist = 1.483*np.median(abs(Z - Z_median))
```

Select the convergence criterion. This choice is important whereas, another value may modify the convergence speed.

```
tol = 1e-5
```

The initial value of X, $s_A^*$ and the number of iterations is setup to 0

```
X_robust = 0
sA_rob = 0
temp_val = 0
nb_iterations = 0
```

Start iteration loop. If the difference between the old median of the $Z_i$ and the new one is above the convergence criterion the loop is pursued.

```
while (abs(Z_median - temp_val) > tol):
```

Update correction criterion. It is used to compute the lower or upper distance from the median that is acceptable. Once more, by simulation it was demonstrated that it must be multiplied by 1.5.

```
  phi = 1.5*max_dist
```

*(Continued)*

At each step, each $Z_i$ value is checked. If it too far above or below the median it is replaced by the acceptable extremum. In that way a new set of $Z_i$ is generated.

```
For I in range(len(Z)):
```

The values that are too low or too high are not deleted but replaced by new ones.

```
if Z[i] < Z_median - phi:
    Z[i] = Z_median - phi
elif Z[i] > Z_median + phi:
    Z[i] = Z_median + phi
```

Update the robust reference value with the mean of modified dataset.

```
X_robust = np.mean(Z)
```

Update the robust standard deviation with the standard deviation of the modified dataset

```
sA_rob = 1.134* np.std(Z)
```

Swap old values with new and increment the number of iterations.

```
Temp_val = Z_median
Z_median = X_robust
max_dist = sA_rob
nb_iterations = nb_iterations + 1
```

Exit of the loop with printing the robust parameters

```
print("Robust")
X_uncertainty = 1.23*sA_star/np.sqrt(len(Z))
print(X_robust, sA_rob, X_uncertainty)
```

Table 4.6 presents the values obtained for the assigned value $X$, its standard-uncertainty, and the standard-deviation of the test $s_A$. Given the type of material analyzed, prepared from a natural sample of unknown content, the most suitable mode of expression for the assigned value, its uncertainty and the standard deviation of the test is given by Algorithm A. For comparison, the values obtained by the classic calculation method are also shown. Table 4.6 shows the role of Algorithm A in modifying the reference value and the standard deviation of the PTS. The major consequence is the reduction of the standard deviation. Both these statistics are used to obtain the scores of the participants and to evaluate their competence. Using the data in Table 4.5, two $z$-scores are computed, one with the classic parameters, the other with the parameters provided by algorithm A.

Figure 4.3 illustrates the results. The legend gives the meaning of each symbol. Individually measured moisture values appear as circles with the axis on the left. Mean values are reported as assigned references. They are slightly different because they are computed by two methods, classic and robust. The score values are read on the right axis. The red diamonds are used for the classic $z$-scores, and the green squares for the robust ones. For ease of interpretation of the scores, two horizontal acceptance limits are added at $\pm 3.0$ units (right axis). Classic scores are more tolerant

**Table 4.6** Assigned value, standard deviation and uncertainty calculated by different methods.

| Method | Reference value $X$ | Standard uncertainty $u(X)$ | Standard deviation $s_A$ |
|---|---|---|---|
| Classic | 8.12528 | Not applicable | 0.19622 |
| Algorithm A | 8.14512 | 0.03271 | 0.15956 |

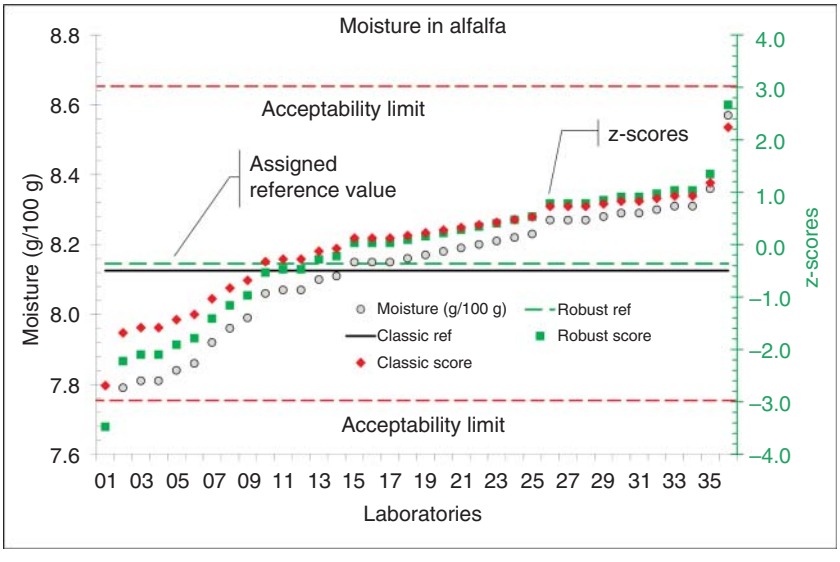



**Figure 4.3** ALFALFA – *z*-scores are reported on the right axis and computed with classic (red diamonds) or robust (green squares) methods. Gray circles are moisture results (left axis).

than robust ones, as no laboratory would be marked as unacceptable, while it is not for one participant with the robust score. The use of PTS results to estimate test sample MU is explained in Section 7.4.1 after the presentation of the general procedure recommended in the GUM.

### 4.3.6 Check Material Homogeneity or Stability

It is obvious that the material selected for a PTS must be homogenous. Considering the regulatory importance of the latter, a laboratory must not get a bad score because it received an imperfect sample. It may therefore seem paradoxical that ISO 13528 proposes to perform homogeneity checking *after* the completion of the test rather than before. The reason given is that it is first necessary to have the reference value of the material $X$ and the standard deviation of the test $s_A$ to apply the recommended statistical control method for homogeneity checking. Homogeneity can be a difficult goal to achieve. Thus, depending on the type of matrix and the nature of the

analyte, it can be difficult to prepare and maintain an homogeneous test material. Homogeneity checking is often achieved at the end of the manufacturing process, while all test samples are ready to be sent. Considering the overall cost of a PTS, test material manufacturing is the most expensive step. For the organizer, finding out afterwards that the test material is not homogeneous remains the worst situation. On the other hand, for the participants, the lack of homogeneity is a real nuisance because the score they obtained may be unacceptable to an auditor. It is therefore understandable how much care the organizer must take in preparing the material and how crucial it is to check it. To do so, ISO 13528 [5] describes a standardized procedure summarized as follows:

1. Choose a laboratory to conduct the homogeneity check measures and the analytical method to use. If it is acceptable not to perform homogeneity check for each required measurand, select the characteristic or characteristics of the material that are most sensitive to heterogeneity, such as moisture in a food.
2. Prepare and package the proficiency test items for a round of the proficiency testing scheme, ensuring that there are sufficient items for the participants in the scheme and for the homogeneity check.
3. Randomly select $I$ test samples, with $I \geq 10$, in their final packaged form.
4. Prepare $J$ test portions, with $J \geq 2$ from each test item using appropriate techniques to minimize between-test-portion differences.
5. Taking the $IJ$ test portions in a random order, obtain a measurement result $Z_{ij}$ (with $1 \leq i \leq I$ and $1 \leq j \leq J$) on each, completing the whole series of measurements under repeatability conditions.
6. Apply a one-way random analysis of variance (ANOVA) as shown in Section 3.2 and Resource E to calculate the grand mean $\overline{\overline{Z}}$, within-samples standard deviation $s_W$, and between-samples standard deviation $s_B$.

The interpretation of the standard deviation depends on the organization of the PTS. But one simple rule is to compare the between-samples standard deviation $s_B$ to the test standard deviation $s_A$ obtained at the end of the PTS. It can be concluded that the test material is homogeneous if:
Acceptance criterion for the test standard deviation

$$s_B \leq 0.3 \times s_A \tag{4.6}$$

Acceptance criterion for the test variance

$$s_B^2 \leq 0.09 \times s_A^2 \tag{4.7}$$

If the variance is considered rather than the standard deviation, this means that the between-samples variance must be less than 9% of the test variance. This is a rule of thumb, and more accurate statistical tests could be applied. In Section 8.3.2, about sampling uncertainty, it is also proposed to use the same experimental design, recommended in an international standard. Computation is simple to obtain for both variances involved in this criterion, as it can easily be achieved with the statistical tools described in Section 3.2.1. Homogeneity checking is a costly requirement since it relies on many measurements. A less expensive method has been proposed

that is based on a non-destructive analytical technique, identified as *rapid* such as near infrared spectrometry (NIRS). NIRS allows a multidimensional approach to homogeneity control, and its ability to highlight small changes in composition that can take place during the manufacturing process is remarkably effective. For this purpose, the entire NIRS spectrum is directly used without having to go through a prediction stage for the constituents [6, 7]. It is also possible to use other statistical techniques based on time series studies, where samples are collected along the manufacturing process. One method of choice is a time series analysis technique called auto-regressive moving average (ARIMA), described by Box and Jenkins [8]. The calculation involved in this method is cumbersome which explains why it was not chosen, although it is extremely sensitive to the smallest variations and allows the detection of slight changes in the spatial distribution of the analyte during the production process.

## 4.4 Control Charts

Another method used to verify the trueness of measurement results is the implementation of control charts. A very convenient way is to use an RM, certified or not, to develop the control chart. Section 7.3 gives more detail on the distinct types of control charts applicable to controlling a production process. The need for a QC system, such as a control chart, can be simple if the analytical method is accurately developed. As stated in ISO 9000 standards, it is not enough to achieve quality, it is also necessary to have procedures to maintain quality over time. Control charts are adequate tools for the purpose. They are conceptually simple: an RM of known concentration $T$ (for Target) is regularly analyzed. Analyzed test portions are often called quality controls or QCs. The measurement values obtained are plotted, in the chronological order of their collection, on a graph called control chart. If non-random variations around the reference value are visually detected, it is assumed that the measurement system is disturbed, and a corrective action must be taken. To simplify detection, two control limits are drawn on each side of the reference value. It is expected that the responses will remain within these limits. In addition, several rules are defined to decide whether the distribution is no longer random and what corrective action is to be implemented to return to a normal situation. Figure 4.4 provides an example of a control chart for Kjeldahl nitrogen determination in wheat flour. The center line represents the reference target value $T$ of the IRM used for this chart. The outer dashed lines symbolize lower and upper control limits *LCL* and *UCL,* and two other dashed lines symbolize lower and upper warning limits *LWL* and *UWL*. Warning lines are not always added to the chart.

This example illustrates how an analytical method can diverge from the statistical process control (SPC) required in quality standards. Three typical nonconforming situations are identified. They illustrate the need to take predefined corrective actions, such as recalibration of the measurement system, which is symbolized by the breaks in the curve in the graph:

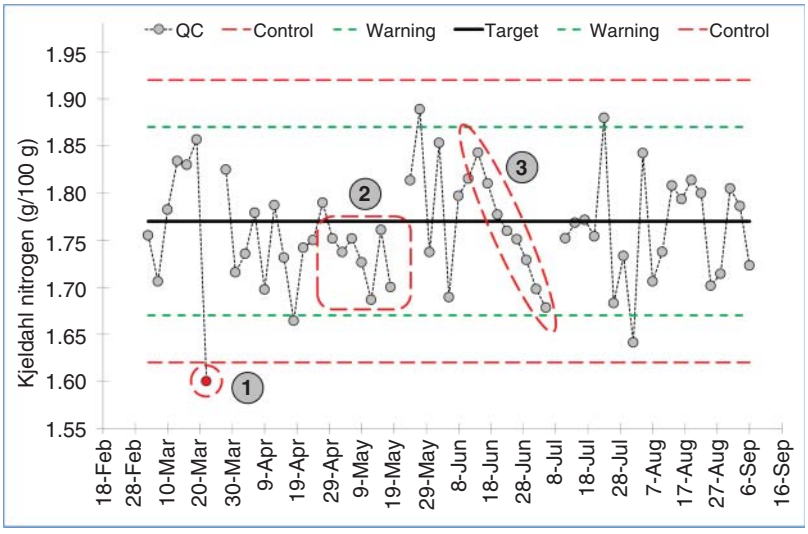



**Figure 4.4** NITROGEN – examples of anomalies in a wheat flour control chart.

① One QC is outside one control limit.
② Seven successive QC are systematically above or below the target value, and as it is expected, they are randomly spread on both sides.
③ Eight successive QCs indicate a systematic upward or downward trend.

Moreover, if any other QC distribution suggests nonrandom behavior, the quality controller is free to take corrective action. The assumption underlying the control charts is that measurements made under statistical control should be distributed according to the Normal law. Control limits can be compared to outlier rejection limits, and it is assumed that any exceeding data corresponds to a malfunction in the measurement process. From a practical point of view, the control chart set up in a laboratory may be different from the industrial context. The easiest way is to use a CRM or an ERM, but their availability or their cost may make this strategy inapplicable. If the laboratory decides to use an IRM, the control chart set up consists of two phases.

### 4.4.1 First Phase Assessment of the Reference Value

1. Prepare homogenized material, in sufficient quantity to complete several tens of test portions. It is best to package each test sample in a sealed bag so that the composition of the material remains stable over time.
2. Analyze at least 30 test portions ($I \geq 30$) with the method to be controlled. When a QC is subsequently planned to involve an average of $J$ replicates, all 30 analyses must be done under the same conditions.
3. Calculate the mean $\overline{Z}$, the standard deviation $\hat{\sigma}$, and the warning ($WL$) and control ($CL$) limits, as described in Eqs. (4.8) and (4.9). Then draw the control chart. The notation $\hat{\sigma}$ is used in this context rather than $s$ because it is the traditional notation for control chart.

### 4.4.2 Second Phase Routine Use

1. Define a control frequency (one per day, for example) and analyze a different QC at each control, the answer is denoted $X_i$. If several test portions are analyzed for each QC, individual values are reported unless a result is expressed as the average of replicates. Section 8.4.3 gives more explanation.
2. When about 20 test portions remain, prepare another internal standard material, measure its contents to define its new reference value and its new standard deviation, and ensure continuity and traceability between old and new control charts.

The standard deviation $\hat{\sigma}$ and the mean $T$ are obtained with the classic following formulas:
Target value

$$T = \overline{Z} = \frac{\sum_i Z_i}{I}$$

Standard deviation

$$\hat{\sigma} = \sqrt{\frac{\sum_i (Z_i - \overline{Z})^2}{I - 1}}$$

Control limits

$$LCL = T - 3\hat{\sigma} \qquad UCL = T + 3\hat{\sigma} \tag{4.8}$$

Warning limits

$$LWL = T - 2\hat{\sigma} \qquad UWL = T + 2\hat{\sigma} \tag{4.9}$$

If each QC corresponds to the average of $J$ replicates, as usually recommended, $WL$ and $CL$ must be differently calculated. New formulas consider this improvement of the control:
Control limits

$$LCL = T - \frac{3\hat{\sigma}}{\sqrt{J}} \qquad UCL = T + \frac{3\hat{\sigma}}{\sqrt{J}}$$

Warning limits

$$LWL = T - \frac{2\hat{\sigma}}{\sqrt{J}} \qquad UWL = T + \frac{2\hat{\sigma}}{\sqrt{J}}$$

For instance, the ISO 7870-2 standard proposes a set of tables that allow the calculation of the limits where measures are expected to lie for Shewhart control charts [9]. Control charts with other distribution laws than the Normal law remain applicable when dealing with counts, such as microbiological methods. It is obvious that a control based based on replicates is more efficient but also more expensive. It is reasonable when measurements are quick and simple, such as physical measurements, temperature, or weighting.

In Section 7.3, a method is proposed to directly derive control charts from the statistical tolerance intervals used to build the MAP. The different limits are obtained

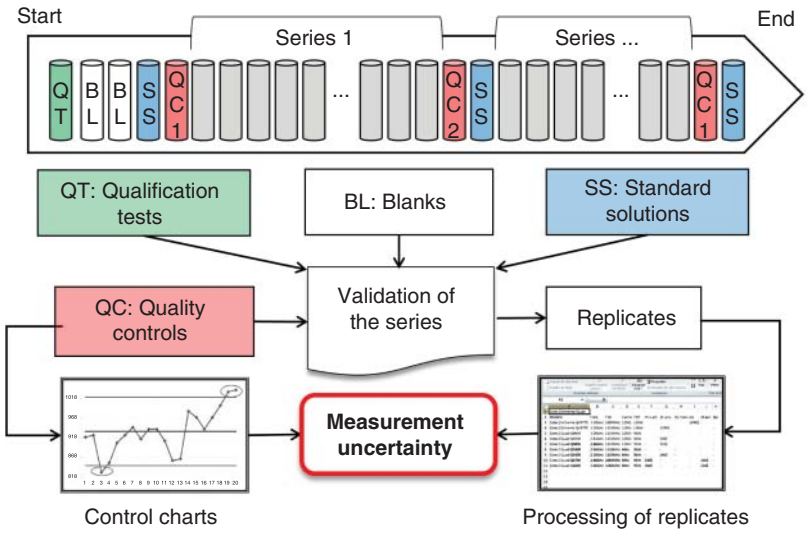



**Figure 4.5** Possible location of different Qcontrol charts in a routine laboratory. Legend: QT, qualification test; BL, blank; SS, standard solution; QC1, QC2, quality controls.

using different values of the probabilities applied to define these intervals. Finally, control charts are a direct application of the principles of theoretical distribution laws. They are increasingly required by auditors of accreditation bodies, as well as participation in proficiency testing. Figure 4.5 shows a possible organization of different control charts for QC in a routine laboratory. The samples sent by the clients to the laboratory are grouped into series and surrounded by diverse types of controls, such as standard solutions, noted SE, blanks BL and, of course, RMs of known content, noted QC1, QC2, etc. Separate control charts can be established for each type of control. At the start of the production line, qualification tests (QT) are applied to the instruments. These tests can also include measurements on commercial reference solutions, as is recommended for many methods, for example, mass spectrometry. The fixed costs for such an organization are not negligible and substantiate the grouping of analyses as series. The remaining question is to define the frequency of the QC.

## References

1 Feinberg, M., Bugner, E., Theiller, G. et al. (1995). Expression of the reference value for proficiency tests. *Journal of Chemometrics* 9: 197–209.
2 Huber, P.J. (2009). *Robust Statistics*. New York: Wiley.
3 Uhlig, S. and Lischer, P. (1998). Statistically-based performance characteristics in laboratory performance studies. *Analyst* 123: 167–172.
4 Analytical Methods Committee of the Royal Society of Chemistry (AMC) (1989). Robust statistics: how not to reject outliers: part 2. Interlaboratory trials. *Analyst* 114: 1699–1702.

5 ISO 13528:2015 *Statistical Methods for use in Proficiency Testing by Interlaboratory Comparisons*. Genève: ISO.
6 Lafargue, M.E., Feinberg, M., Daudin, J.J., and Rutledge, D.N. (2003). Detection of heterogeneous wheat samples using near infrared spectroscopy. *Journal of Near Infrared Spectroscopy* 11 (2): 109–121.
7 Lafargue, M.E., Feinberg, M., Daudin, J.J., and Rutledge, D.N. (2003). Homogeneity check of agricultural and food industries samples using near infrared spectroscopy. *Analytical and Bioanalytical Chemistry* 375 (4): 496–504.
8 Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day.
9 Standard ISO 7870-2:(2023) (2023). *Control Charts — Part 2: Shewhart Control Charts*. Genève: ISO.

# 5

# Method Validation

## 5.1 Review of Validation Procedures

For many decades, analysts applied analytical methods developed in their laboratories. They rarely fully documented the operating procedure, as well as the tips and tricks related to their own experience, and often preferred an oral transmission of their knowledge. Obviously, they knew that random and systematic errors due to matrix interferences or incorrect manipulations might occur, but they also believed that their expertise was good enough to detect them and guarantee the correct interpretation of the result. This state of mind ended in the fifties for the Pharmaceutical community with the tragic problem of Thalidomide and the mid-1970s when various industrial accidents occurred (Minamata, Bhopal, Seveso, etc.).

It led end-users, especially control authorities in charge of public health or the environment, comparing laboratory results with others, to surprisingly realize that extremely high differences could be observed. For instance, in the 1980s, differences as high as to 1/1000 could be observed when analyzing chlorinated pesticides in foods. Motivated by the expectations of their customers, analysts set about two complementary tasks: harmonize the operating procedures and validate the methods.

For the first goal, working groups were established within national and international standardization organizations, such as ISO, or professional structures, such as AOAC. This resulted in many documents being presented in various forms. It is beyond our scope to describe the work that has been done. However, the fundamental role of standardization and harmonization efforts in improving the metrological quality of methods must be underlined and encouraged.

Here again, it is often through collaborative work that the second objective was achieved. Two complementary tasks can be distinguished:

- Develop the vocabulary with definitions of various performance characteristics or figures of merit.
- Develop the validation procedures with respect to expected performance criteria.

Despite the generic aspect of this program – since all analytical methods require validation regardless of their field of application – it has been implemented by a wide variety of organizations, resulting in various proposals and sometimes confusion.

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

The emergence of the need for validation of analytical methods can be dated back to the late 1980s.

This period overlaps with the publication by ISO of the first international standards on quality assurance: the ISO 9000 series. In these documents, many concepts related to "quality and all operations that contribute to it" are defined. Validation can be considered as a procedure to reach the adequate quality level for an analytical measurement. In the meantime, it was classically claimed that the quality of products or services must be designed to "satisfying customer needs." This is also known as a "consumer-driven quality," which was summarized in a five-point list when applicable to the laboratory:

- Meet a well-defined need, use, or objective.
- Fulfill consumer expectations.
- Comply with standards, specifications, regulations, and other requirements.
- Have at a competitive price.
- Provide data at a cost that generates a profit.

Considering a more recent context, new requirements should be added:

- Reduce the environmental impact of laboratories.
- Avoid reagents and substances unsafe for personnel and the environment.

From the beginning, the application of quality assurance principles in the laboratory has been perceived as a specific objective that required as such, specific texts. Today, the list is long of guides, recommendations, and guidelines published by various regulatory or professional structures on the application of quality assurance in the laboratory and on method validation. For example, a large set of official documents available on the Internet describes all operations required to fulfill quality objectives [1–6]. Different comparative reviews of these documents have been published [7, 8].

Validation *per se* of methods often fills a short section of these documents, but it is always addressed. In most guides, the principle to conduct the validation consists in collecting experimental data that are processed to assess performance characteristics, such as accuracy, precision, limit of detection and limit of quantification, or sensitivity. Sometimes, schemes are also prescribed for the practical organization of the experiments.

But, except for some guidances, little or no advice is given on the number of measurements to be collected for a correct estimation of the performance characteristic nor on the experimental design to be applied. The next Sections 5.4.2 and 5.4.3 explain how the appropriate number of replicates is rarely used and may impact the reliability of obtained results and how a correct design of experiment can be crucial. Consequently, the decision rule for concluding whether a method is valid or not is not always very conclusive.

Over the last decades, validation has become an important concern for many analysts, and many of them have acquired obvious expertise. However, it may be useful to clarify some incoherent points, such as the vocabulary and the structure of the experimental design, where ambiguities persist. The imprecision of the classic

validation guides led us to develop a more structured approach. A statistical and graphical tool for method validation, called method accuracy profile (MAP), resulted from this observation.

A clear interpretation strategy is also in the pipeline. Over the past two decades, numerous papers using the accuracy profile have been published, as explained in this chapter. For several official standardization bodies, *validation* is synonymous with interlaboratory study. But, in many situations, method validation is internally conducted in a single laboratory and is sometimes called in-house validation. For example, if the method is:

- Developed for a time limited research topic.
- Applied to a small series of samples.
- Used to control the manufacture of a single product by a single producer.
- Transferred from another laboratory and its reliability needs to be verified.
- Restarted after a period of inactivity.
- The property of a laboratory that does not wish to transmit it to others.
- Multianalyte and multimatrix, for which an interlaboratory study is difficult to organize.

For most of these cases, it is not necessary or possible to involve a number of laboratories to validate a method. More generally, interlaboratory validation applies to a method in a pre-competitive context, such as health or official control, whereas an in-house validation applied to a method used in a competitive context, such as drug control.

### 5.1.1 Inconsistencies of Validation Vocabulary

While validation is a common requirement for all analysts for all methods, it is paradoxical that the documents related to this topic are generally presented as specific to an application sector such as drugs, foods, metals, medical biology, oenology, water, or mining. It is even more surprising because laboratories, whatever they work on, use the same analytical instruments and have the same validation needs. Among them, the main differences are:

- In the pre-analytical sampling step and sample preparation,
- In the final validation criterion.

As detailed in Section 8.4.3 about the definition of a replicate, only the instrumental step is purely *analytical* as it requires measuring principles based on chemistry or biology. The economic evolution of laboratories over the last few decades has pushed them to get together within larger industrial organizations and to be much more flexible than in the past.

For example, some years ago, soil and water analyses were conducted in specialized and independent laboratories. Today, it is common for these types of activities to be associated within a single legal structure and, sometimes, even in the same location. Because of the classic sectorial or *vertical* organization of many application fields, a diverse list of definitions has been independently issued, raising several communication problems.

To avoid some of these ambiguities, the reference proposed in this book is the International Vocabulary of Metrology or VIM, published by the Bureau International des Poids et Mesures or BIPM [9]. It is a well-accepted thesaurus applied in various fields of measurement. The definitions of many terms, when they exist, apply to analytical sciences, although some are missing or unsuitable. Most of the definitions are quoted in the text and written between double quotes. As a remainder, in VIM, two major vocabulary elements can be identified: characteristics and parameters.

| | |
|---|---|
| Characteristics | They are *concepts* not "directly measurable" such as trueness, precision, or accuracy. For example, note 1, coming with the definition of measurement accuracy (VIM clause 2.13), states that: "measurement accuracy is not a quantity and is not expressed numerically. A measurement is sometimes said to be more accurate if it provides a smaller measurement error." |
| Parameters | They are estimated from experimental data using a computational algorithm and called *estimators* by statisticians, i.e. a rule for calculating an estimate of a given quantity based on observed data. They are the coefficients of mathematical or statistical models. |

For example note 1 on measurement precision (VIM clause 2.15) states that: "precision is usually expressed numerically by parameters such as standard deviation, variance or the coefficient of variation under the specified conditions." If one parameter gives a numerical evaluation of a characteristic, then one characteristic can have several parameters alone or in combination. The numerical values of the parameters are sometimes referred to as *performance scores* or *figures-of-merit*.

A third category of vocabulary elements must be added, absent from the VIM, but essential to validate a method.

| | |
|---|---|
| Criteria | They are *technical specifications* expressed as a minimum, maximum, acceptance interval, etc. They are often quantitative but sometimes qualitative, such as practicability or robustness. The procedural manual of the Codex Alimentarius Commission (CAC) [10] by means of the Codex Committee on Methods of Analysis and Sampling (CCMAS), proposes a Criteria Approach for selecting methods of analysis for food trade that is based on a set of criteria. |

A "Glossary of Used Terms" is proposed as an attempt to sum up some the usual definitions. It results from the compilation of various documents, including the VIM. It is not intended to be exhaustive or referential, as several instrumental methods or sectors of activity have developed their own vocabulary, but only to collate the definitions of the main terms used in this book.

Given the inconsistency of glossaries, it is not surprising that several authors underlined the divergences in published documents. The same term can have different definitions and meanings, sometimes difficult to relate to. The term "method validation" illustrates these discripancies, as it can mean:

- "Verification, where the specified requirements are adequate for an intended use" (VIM), but also.
- Interlaboratory comparison that leads to precision parameters, such as reproducibility.

– Demonstration that the method measures the analyte it is supposed to. This definition is used in the title of many publications as a synonym for specificity.

The term "linearity" also has different meanings:

– Proportionality between the known contents of reference material and the inverse-predicted concentration, i.e. linearity of trueness.
– Existence of a linear relationship between instrumental response and analyte concentration, i.e. linearity of the calibration.

For the limit of quantification (LOQ), dozens of definitions are available, and even more calculation methods. Yet it is the most widely used parameter by analysts or equipment manufacturers to advertise the performance of an analytical technique. This topic is discussed in more detail in Section 9.1.

Another critical issue is the eventual confusion between the terms applicable to quantitative methods with those used for qualitative tests, as their concepts are different. A compilation of nearly forty validation guides, published mainly by international organizations, identified various inconsistencies. Five benchmarks were selected to conduct this textual analysis [8]. They are concerned with the type of analysis, chemical or biological, the national or international scope of the guide, the sector or discipline of origin, the analytical technique, and the compounds analyzed.

The various validation guides share a common set of terms, despite variable definitions, as shown in Figure 5.1. Perhaps it is this appearance of similarity that currently prevents a global standardization effort from taking shape.

### 5.1.2 Validation Plans

When scrutinizing literature, it is possible to identify a *classic validation procedure* that is widely published. IUPAC directive or ICH Q2(R2) guidelines are typical examples of such a procedure [11, 12]. The content of a classic validation procedure is as follows:

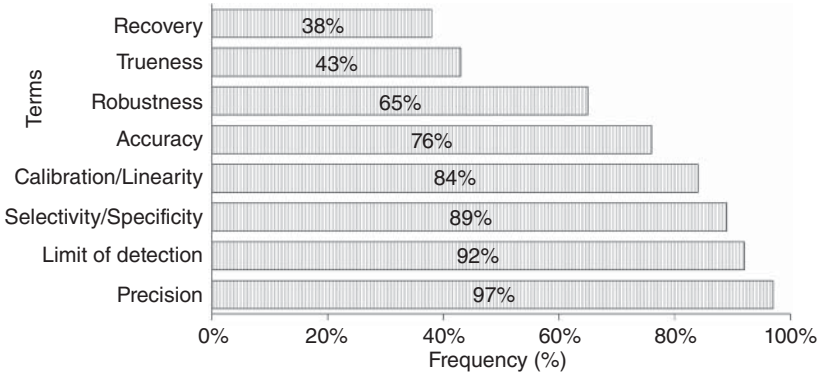



**Figure 5.1** Frequency of terms used in validation guides. Source: Adapted from Raposo and Ibelli-Bianco [8].

- A list of characteristics to be addressed in the validation experiment, such as precision, accuracy, applicability, very often including selectivity, response function and calibration, etc.
- One or more parameters for each of these characteristics are aimed for providing quantitative estimates, such as standard deviation, recovery rate, and percentages.

If no acceptance criteria are explicitly set for evaluating the estimated values of the parameters, the analyst only performs a characterization of the method, which is not a full validation. To be complete, the values of the parameters must be checked as conforming to performance criteria values. In practice, the statistical approach proposed to estimate parameters is simple numeric calculations, such as average, standard deviation, or the coefficient estimators of a function obtained by using the least-squares regression method. For qualitative methods, it is a matter of proportions, with confidence intervals obtained by referring, most often, to a binomial distribution. Therefore, all these calculations can easily be done with simple software, such as Microsoft Excel or Open Office Calc.

A more recent trend is to apply more elaborate algorithms, such as the weighted least-squares (WLS) method or partial least-squares (PLS) regression (Section 2.3). These methods require computation procedures unavailable in worksheets. In many analytical devices, such as mass spectrometers or near infra-red spectrometers, manufacturers have included them directly in the monitoring software. But, the Python language can help the analysts needing to develop their solutions.

The diverse estimators proposed in different validation guides to estimate the same parameter is another confusing issue for analysts, such as robust algorithms to estimate reproducibility [13]. Other computational methods requiring intensive calculation, such as Monte Carlo simulation, remain in the domain of specialized scientific publications in analytical sciences but are usual in other measurement fields. This specific topic is not addressed in this book.

Sometimes statistical tests, known as significance tests, are recommended to assess the conformity of a parameter to a criterion [14]. Unfortunately, only the first type of risk of error, denoted $\alpha$ risk, is usually considered. This is also known as the producer's risk or the risk for an analyst who used a valid method but concludes that it was not. In contrast, the customer's risk, denoted $\beta$-risk, i.e. the risk to the client that the method is not valid but regarded as valid by the analyst, is not assessed.

It is exceptional that the validation procedure indicates how to organize the experimental design. This is left to the experimenter's initiative, who may proceed randomly and intuitively. This topic is discussed in more detail in Section 5.4.2. While a list of characteristics and/or parameters to be experimentally estimated is required, the analyst remains free to choose the experimental design that is considered the most appropriate to achieve this goal.

The classic validation procedure can be called a multicriteria method because it consists of evaluating several characteristics separately compared to a set of criteria and making individual decisions on each. Therefore, contradictory conclusions may occur, such as conforming precision but non-conforming trueness. The final decision is difficult to make. While a list of characteristics and/or parameters to be

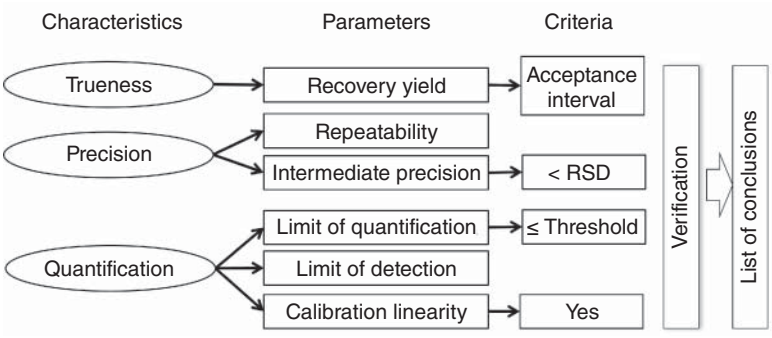



**Figure 5.2** Example of a multicriteria validation procedure.

experimentally estimated is required, the analyst is free to choose the experimental design that is considered the most appropriate to achieve this goal.

As illustrated in Figure 5.2, based on the characteristics and parameters required in ISO 17025 standard, the classical validation strategy can be described as a *multicriteria* method because it consists of evaluating the different characteristics sequentially and separately against a set of criteria and making an individual decision on each one [15].

This strategy may provide a lot of interesting information to the analyst, but practice shows that it has the disadvantage of leading to a set of conclusions that may be contradictory. For example, precision is conforming, while trueness is not. How to conclude? The final decision may quickly become subjective. But the major downside is that it is not evident that this knowledge is the proper answer to "satisfying customer needs" as claimed in quality assurance standards.

For instance, US-FDA has set up a list of acceptance criteria to define a validated method called standard method performance requirement or SMPR [16]. This document has a set of criteria values for precision and recovery yield. These values are included in several other official documents issued by different institutions, such as the European Commission or the Codex Alimentarius [1, 10].

But, when comparing several guides applicable at the international level for regulatory or health purposes, it is surprising that in many cases, no acceptance criteria nor verification procedures are proposed. The number of degrees of freedom available for decision-making is generally ignored. However, it is crucial when statistical testing is required, as underlined in many statistical handbooks.

Alternatively to the classic strategy, it is possible to propose a validation procedure based on a unique criterion. The latter amounts to choosing a unique comprehensive characteristic which may be a combination of several parameters. Then by comparing this parameter combination to a single criterion, it is possible to obtain more easily a final decision. Such single-criterion validation procedure are less often used. A possible example is using the total analytical error (*TAE*) described below.

But it is also appropriate to select accuracy as a unique validation characteristic, which is defined as the "closeness of agreement between a measured value and a true

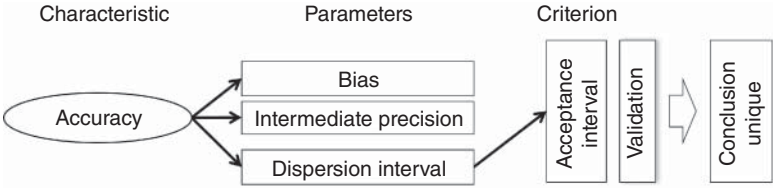



**Figure 5.3** Example of a single-criterion validation procedure.

value of a measurand" and aims to verify if this closeness is satisfactory. A detailed discussion about the concept of accuracy in this context is given in Section 6.10.

To mathematically express accuracy, it is promising to combine several parameters into one statistical parameter, for instance, one so-called "statistical dispersion interval." This lesser-known family of statistical intervals is described in Section 5.3. The latter is compared to an acceptance interval, resulting in a unique parameter *versus* a unique criterion.

The MAP is a validation procedure described in Section 5.2 that is an example of a single-criterion strategy based on accuracy, slightly different from *TAE*. It was developed in the late 1990s, in the framework of a commission of the Société française des sciences et techniques pharmaceutiques (SFSTP). It was published in a series of papers [17–20]. Figure 5.3 illustrates schematically this single-criterion validation procedure.

This method, applied for nearly 20 years in many laboratories, on a wide range of methods, and in varied scientific contexts (Section 5.1), was adopted by several international standards addressing method validation, such as ISO-16140 or ISO 22116.

MAP was chosen to describe this procedure and accuracy as a unique characteristic because it is often identified as the combination of trueness and precision (even the correct "combination" to be used is undefined). For example the ISO 5725 standards have the general title: "Accuracy (trueness and precision) of measurement results and methods." It seemed logical to estimate accuracy by combining two parameters, the first for trueness, such as a bias or a recovery rate, and the second for precision, such as a standard deviation of precision.

A comparable combination was already proposed in the 1970s to introduce the concept of total analytical error, or *TAE,* thoroughly described in the first Section 6.1. of Chapter 6 [21]. The recently revised definition of accuracy [22] explains that it is a "non-numerical" concept. The initial choice seems inconsistent. Nevertheless, for convenience and because it is now well-known, the term method accuracy profile is retained here. There is also a possible confusion with the use of accuracy in some documents: accuracy is not trueness as explained in Chapter 4, although both definitions seem relatively close, as visible in the glossary at the end of the book (see Glossary of Used Terms).

The estimation of the precision is achieved by means of a standard deviation. As explained in Section 2.4.3, ISO 5725 standards define at least three experimental conditions for computing a precision standard deviation, namely reproducibility, repeatability, and in-between intermediate precision. The reproducibility

condition is considered well-adapted to interlaboratory validation studies. But the intermediate precision condition is most adequate for in-house validation because it allows accounting for various sources of laboratory-specific variation.

As described in Section 7.2, intermediate precision is best adapted to a simple and direct estimation of a large part of measurement uncertainty (MU). It is recognized that this goal is out of reach under repeatability conditions. With hindsight and experience, when comparing classic procedure and MAP, some major advantages can be stressed, as MAP permits:

- To compute statistical dispersion intervals (Section 5.3) that give an insight into the measurement value scattering and to verify that it complies with the method acceptance performance.
- To yield a comprehensive procedure in estimating MU as justified in Section 7.2.
- To apply a consistent experimental design which helps verify if the number of trials is optimized as discussed in Section 8.4.3.
- Finally generate the method uncertainty function applicable to any future sample analyzed at the laboratory as explained in Section 7.5.

For these reasons, the MAP is used as a roadmap throughout the whole book and presented as an adequate solution to many problems the laboratory faces for validating a method and finally to have access to a relatively simple estimation procedure of MU.

## 5.2 Method Accuracy Profile (MAP)

### 5.2.1 Principles

Because incorrect measurements can lead to wrong decisions and, therefore, to significant additional costs or even risks, MAP is a validation procedure founded on the following reasoning: to validate a method means to demonstrate that it can produce an important, even prerequisite, proportion of measurement values serviceable for a correct decision, for example at least 80% or 90% of possible measures are distributed within an acceptance interval around the true value of a sample. Whereas accuracy is defined as the "closeness of agreement between a measured value and a true value of a measurand" this parameter can generate confidence in the analytical result, once estimated.

The acceptance interval is specified as $[A_L; A_U]$, preferably pre-defined by the analysis end-user. The subscript $L$ of $A_L$ stands for lower bound, and $U$ for upper bound. In other words, the analyst can claim the method as validated, with a known confidence level, when it is able to produce at least the required proportion of data lying within the acceptance interval bounds. The major innovation of the suggested method is to use statistical dispersion intervals to compute how much acceptable measures are scattered. These lesser known statistics are estimators of the probable dispersion of results and are expressed as intervals.

Ultimately, this approach is more straightforward than computing separate statistical parameters, such as means or variances, and independently checking

whether they conform. Consequently, in statistical terms, to demonstrate validity and check whether the method fulfills defined requirements, the proposed approach is to calculate the statistical dispersion interval within which lies the required proportion of measurements and verify it is included within the acceptance interval. As explained below, several types of statistical dispersion intervals exist and can be applied to this goal. Starting from this reasoning, the MAP harmonized procedure was established and published between 2004 and 2008 in the *Journal of Pharmaceutical and Biomedical Analysis* [17–20].

A mnemonic presentation of MAP consists of the 10-step list registered in The 10-step MAP procedure provided at the end of the book. The importance of the concept of *series* indicated at step 4, was already underlined in Section 3.1 about interlaboratory studies. A series includes all measurements performed under identical conditions, for example, the same method, the same day, the same operator, the same calibration curve, etc. This reflects the usual laboratory management where samples submitted to the same analysis are grouped before getting started. This organization aims to reduce the fixed costs of an analytical operating procedure as explained in Section 4.4 about control charts. Because a method must be validated the way it will be used in the future, the practical layout of the series should reflect as closely as possible the sources of uncertainty likely to occur during routine application of the method. If the measurements are not organized into series, it is impossible to complete the calculations and construct the MAP. The requisite of the experimental design is explained in Section 5.4.2.

In the 1990s, some guidelines were proposed to validate a method for a single concentration level. But it was quickly realized that the figures of merit of a method – precision and trueness – vary with the concentration. Therefore, it is now considered mandatory to perform validation measurements on at least three validation materials of different concentrations spreading over the validation range.

Once measurements are collected and inverse-predicted concentrations are obtained for each material, the bounds of β-expectation tolerance intervals (β-ETI) are calculated by combining the different validation parameters. Statistical dispersion intervals consist of a family of statistical parameters less well-known than confidence intervals, thus extensively explained in Section 5.3.1. They are designed to contain a certain proportion, denoted $\beta$, of future outcomes or sometimes a single future outcome.

The letter $\beta$ is the coverage probability of the interval and corresponds to this proportion. In the text, it is denoted $\beta$ when expressed between 0 and 1, such as 0.80, or $\beta$% as a percentage, such as 80%. These two equivalent notations are a reminder that in worksheets, it must be distinguished between the number, as stored in computer memory, and its display format, as visible in a worksheet cell.

Finally, for a given validation material, the β-ETI corresponds to an interval where it can be predicted that a proportion $\beta$% of future measurement values may fall. By convention, the coverage probability is denoted $\beta$ and should not be confused with the risk of error $\beta$ of hypothesis testing. It can be expressed as:

Generic expression of the statistical dispersion interval

$$[\overline{Z} \pm k_{TI} \times s_{TI}] \tag{5.1}$$

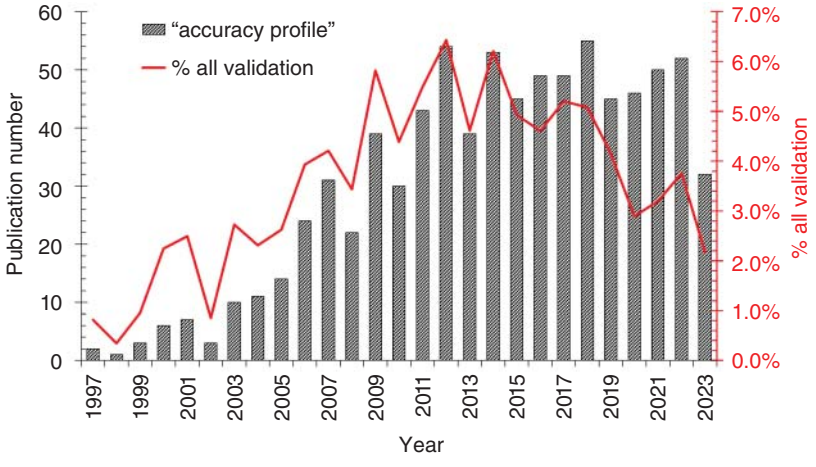



**Figure 5.4** Number of publications with "accuracy profile" in the title ratioed to all "validation" published papers.

where $\overline{Z}$ is the average inverse-predicted concentration of replicates of the validation material, $k_{TI}$ the coverage factor of the interval depending on the chosen probability $\beta$, and $s_{TI}$ the standard deviation, which will be defined further. At each concentration level, i.e. for each validation material, the β-ETI bounds are separately computed and are graphically connected to draw a profile, giving its name to the procedure.

It has been recognized that the performance of a method is greatly dependent on the concentration in a nonlinear way (see Section 7.5). Therefore, the linear interpolation between interval bounds is incorrect, but it is a simple and convenient solution to illustrate and graphically interpret a MAP. But the use of straight-line segments when profiles are expressed as percentages to the reference concentration can be misleading for some complementary calculations as explained in the next section about the limit of quantification.

Since the first publication in 1996, the number of validation examples based on the MAP has increased. Figure 5.4 illustrates the results of a simple query of the document database Science Direct, run by Elsevier. Searched papers must contain the term "accuracy profile" in the title. About 815 publications were found, but it can be assumed that this number is an underestimation since this validation procedure was standardized and may have been used but not cited in the title, or unpublished analytical methods were validated using an accuracy profile.

Over the same period, the terms "method validation" and "analytical method" appeared in approximately 21,000 publications in chemistry journals. Thus, 2% and 5% of the publications were based on the MAP procedure. This significant difference is due to semantic ambiguity. Many analysts use the term "validation" when developing or characterizing a new method and just checking the specificity, i.e. verifying the method is correctly determining the analyte it is supposed to.

**Table 5.1** THEOPHYLLINE – description of the dataset.

| Title of the dataset | THEOPHYLLINE |
|---|---|
| Reference publication | [23] |
| Measurand | Theophylline concentration in plasma, expressed in µg/l |
| Measurement method | Ultra-High Pressure Liquid Chromatography (UHPLC) coupled with a tandem mass spectrometer detection UHPLC-MS/MS |
| Validation range | [0.05, 10] µg/l |
| Acceptance interval | ±25%. Classic value for biological analysis |
| Validation materials | Six validation materials containing 0.05, 0.1, 0.5, 1.0, 2.5, and 10.0 µg/l respectively, prepared by SAM from a batch of homogenized plasma |
| Validation design | Series ($I = 6$), replicates/series ($J = 2$), levels ($K = 6$) |
| Calibration design | Series ($I = 6$), replicates/series ($J' = 2$), levels ($K' = 5$) containing 0.02, 0.1, 0.5, 2.5, and 10.0 µg/l respectively |
| Total number of measures | 60 measurements on calibration solutions. 72 measurements on spiked materials. |
| Predicted concentrations | Table 5.2 |
| Statistics | Table 5.3 |

### 5.2.2 Method Accuracy Profile by Example

The THEOPHYLLINE dataset already used in Section 2.2 will illustrate the steps described above in building the MAP and explain the final interpretation. Details on instrumental conditions and sample preparation are described in the original publication [23]. The dataset features are summarized in Table 5.1. Calibration data are not presented, except for the series 1 data already put together in Table 1.3 to illustrate the different applicable algorithms to estimate a calibration curve. As explained, different calibration models and regression methods apply to the same dataset to obtain different calibration curves and, therefore, inverse-predicted concentrations. For all series, the latter expressed in µg/l, are gathered in Table 5.2. They were calculated using second-order polynomials fitted by WLS regression. In Section 8.1, the role of the different regression techniques will be illustrated to demonstrate how the accuracy profile is a very efficient tool for selecting the best calibration function.

All details on the formulas used to obtain inverse-predicted concentrations are already described in Section 5.3.1. Compared to the original publication [23], the results in Table 5.2 are rounded to facilitate a possible use as benchmark. This remark is intended to explain slight differences between the calculated data and those originally published. Because many formulas use sums of squares, an even small difference may result in an even more important difference in the final result. In Table 5.3, we gather the basic statistical parameters that are necessary to draw

**Table 5.2** THEOPHYLLINE – inverse-predicted concentrations of the validation materials obtained by WLS quadratic regression (µg/l).

| | | Series | | | | | |
|---|---|---|---|---|---|---|---|
| **Calibrator (µg/l)** | **Replicate** | **1** | **2** | **3** | **4** | **5** | **6** |
| 0.05 | 1 | 0.077 | 0.052 | 0.055 | 0.049 | 0.051 | 0.076 |
| | 2 | 0.074 | 0.058 | 0.056 | 0.049 | 0.052 | 0.055 |
| 0.1 | 1 | 0.114 | 0.112 | 0.104 | 0.100 | 0.113 | 0.147 |
| | 2 | 0.113 | 0.110 | 0.101 | 0.105 | 0.107 | 0.112 |
| 0.5 | 1 | 0.534 | 0.509 | 0.479 | 0.593 | 0.538 | 0.506 |
| | 2 | 0.543 | 0.494 | 0.478 | 0.535 | 0.512 | 0.514 |
| 1.0 | 1 | 1.144 | 1.028 | 0.902 | 0.988 | 0.977 | 0.975 |
| | 2 | 1.113 | 0.996 | 0.892 | 1.074 | 0.957 | 0.970 |
| 2.5 | 1 | 2.560 | 2.372 | 3.127 | 2.888 | 2.380 | 2.420 |
| | 2 | 2.486 | 2.233 | 2.280 | 2.585 | 2.394 | 2.472 |
| 10 | 1 | 10.424 | 10.164 | 9.928 | 10.037 | 10.134 | 10.470 |
| | 2 | 10.829 | 10.606 | 9.286 | 10.832 | 10.518 | 10.998 |

a MAP. In this example, $\beta$% is set equal to 80%, which means that each $\beta$-ETI will contain, on average, 80% of the future measurements that could be obtained on these validation materials. For example at the concentration level of 0.5 µg/l, 80% of possible measurement values are expected to lie between [0.470, 0.569] µg/l. Explanation and computation details are given further in Section 5.3.1.

The results summarized in Table 5.3 were obtained using the worksheet program called Resource H β-ETI (Excel) This remark is intended to explain that they can easily be reproduced by any analyst. The interval bounds are not directly used to represent the MAP graphically. Beforehand, they are ratioed to the reference value $X$ assigned to each validation material. This convenient mode of expression is a percentage that can be interpreted as a recovery yield bounded with its dispersion interval β-ETI.

The relative values of TI limits must not be confused with the relative standard deviation (*RSD*) classically used by analysts to express method performances as described in Section 3.2.1. For instance, at level $X = 0.050$ µg/l, the average inverse-predicted concentration is $\overline{Z} = 0.059$ µg/l, and the standard deviation of repeatability $s_r = 0.0064$ µg/l. Thus, the *RSD* of repeatability for this level is $RSD_r = 100 \times \frac{0.00064}{0.059} = 10.8\%$. While the recovery yield interval at the same concentration is completely different [84%; 150%]. *RSD* is not interesting for the present objective of validation because it is not compared to an acceptance criterion. If $\overline{Z}$ was used instead of $X$ to compute the relative TI bounds, the biases would be included and may modify the resulting values. Ratioing the intervals is also a way to illustrate the method profile in a familiar manner for analysts.

**Table 5.3** THEOPHYLLINE – main validation parameters and β-expectation tolerance interval bounds.

| **Parameters** | | **Levels** | | | | | |
|---|---|---|---|---|---|---|---|
| Concentration (µg/l) | $X$ | 0.050 | 0.10 | 0.50 | 1.0 | 2.5 | 10.0 |
| Lower acceptance limit | $A_L$ | 0.038 | 0.075 | 0.375 | 0.750 | 1.875 | 7.500 |
| Upper acceptance limit | $A_U$ | 0.063 | 0.125 | 0.625 | 1.250 | 3.125 | 12.500 |
| Precision parameters | | | | | | | |
| Number of series | $I$ | 6 | 6 | 6 | 6 | 6 | 6 |
| Number of measures | $I \times J$ | 12 | 12 | 12 | 12 | 12 | 12 |
| Number of replicates | $J$ | 2 | 2 | 2 | 2 | 2 | 2 |
| Predicted concentration | $\overline{Z}$ | 0.059 | 0.112 | 0.520 | 1.001 | 2.516 | 10.352 |
| Recovery yield | % | 117 | 112 | 104 | 100 | 101 | 104 |
| Repeatability std. dev. | $s_r$ | 0.0064 | 0.0104 | 0.0192 | 0.0287 | 0.2641 | 0.3905 |
| Between-series std. dev. | $s_B$ | 0.0089 | 0.0067 | 0.0266 | 0.0748 | 0.0000 | 0.2841 |
| Intermediate Precision | $s_{IP}$ | 0.0110 | 0.0124 | 0.0328 | 0.0802 | 0.2641 | 0.4829 |
| β-Expectation Tolerance Intervals (β-ETI) | | | | | | | |
| Tolerance interval std. dev. | $s_{TI}$ | 0.0117 | 0.0130 | 0.0350 | 0.0862 | 0.2749 | 0.5093 |
| Degrees of freedom | $N_E$ | 7.01 | 9.59 | 7.02 | 5.69 | 10.91 | 9.22 |
| Coverage factor | $k_{TI}$ | 1.41 | 1.38 | 1.41 | 1.45 | 1.36 | 1.38 |
| Proportion | $\beta$% | 80% | 80% | 80% | 80% | 80% | 80% |
| Lower bounds | | 0.042 | 0.094 | 0.470 | 0.876 | 2.141 | 9.649 |
| Upper bounds | | 0.075 | 0.129 | 0.569 | 1.126 | 2.891 | 11.055 |

Notations are explained in Section 5.3.1.

Figure 5.5 is based on Table 5.4 data and provides a typical example of an accuracy profile. Since the validation range is widely scattered, covering six concentration levels and three orders of magnitude between 0.05 and 10.0 µg/l, the *funnel* shape of the profile is visible. For the above-explained reasons, relative TIs are not symmetrical in respect to the 100% recovery line. The elements used to draw the MAP are as follows:

On the horizontal axis:

– The theoretical concentration of the analyte, denoted $X$, assigned to the validation materials and obtained, in this example, by standard additions.

On the vertical axis:

– The average recovery yield is the gray dots.
– The two bounds of the β-ETI of inverse-predicted concentrations $Z$, expressed as recovery yields and connected by interpolation solid lines.
– The limits of the acceptance intervals are two horizontal dashed lines, defined according to the objective of the method, also expressed in %, like the β-ETI bounds.

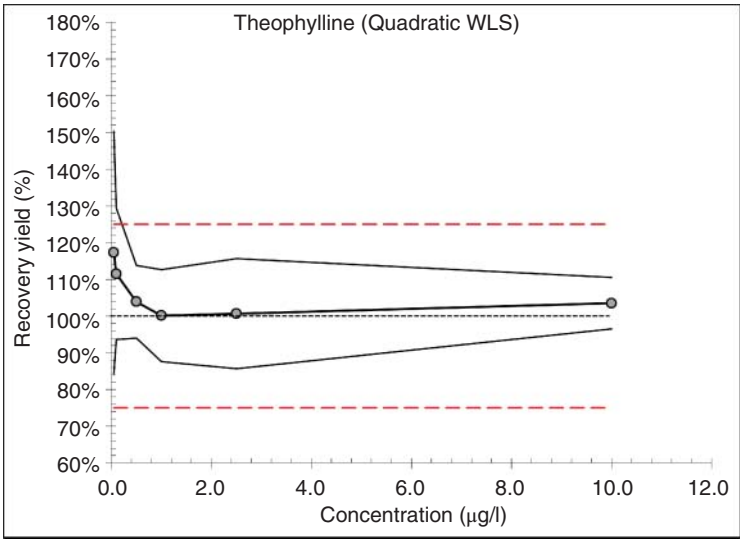



**Figure 5.5** THEOPHYLLINE – MAP with six validation materials. Inverse-predicted concentrations are obtained with WLS quadratic model. Dots: average recovery yields. Solid lines: β-ETI for $\beta\% = 80\%$. Dashed lines: acceptance intervals (±25%).

**Table 5.4** THEOPHYLLINE – β-expectation tolerance intervals relative to the reference values of the corresponding level.

| Levels (µg/l) | 0.050 (%) | 0.10 (%) | 0.50 (%) | 1.0 (%) | 2.5 (%) | 10.0 (%) |
|---|---|---|---|---|---|---|
| Recovery yield (%) | 117 | 112 | 104 | 100 | 101 | 104 |
| Expectation tolerance interval $\beta = 0.80$ | 84 | 94 | 94 | 88 | 86 | 96 |
| | 150 | 129 | 114 | 113 | 116 | 111 |
| Acceptance interval | 75 | 75 | 75 | 75 | 75 | 75 |
| | 125 | 125 | 125 | 125 | 125 | 125 |

The graphical interpretation rules are simple and summarized in Figure 5.6:

1. When the bounds of the β-ETIs are within the acceptance interval, the method is said to be valid at this level of concentration because it can be predicted that at least $\beta\%$ of future measures will lie inside the acceptance interval, for instance at ±25% around the analyte assigned true value in this study.
2. When a single bound is outside the acceptance interval, the method is no longer considered capable of providing the expected proportion of accurate measurements.

It is possible to define the validated range of the method in more statistical terms. It is the concentration range where at least 80% of the measurements the method provides are considered acceptable. In Section 5.4.6, it is explained how to compute the proportion of results exceeding the bounds of the β-ETIs and defined as nonacceptable.

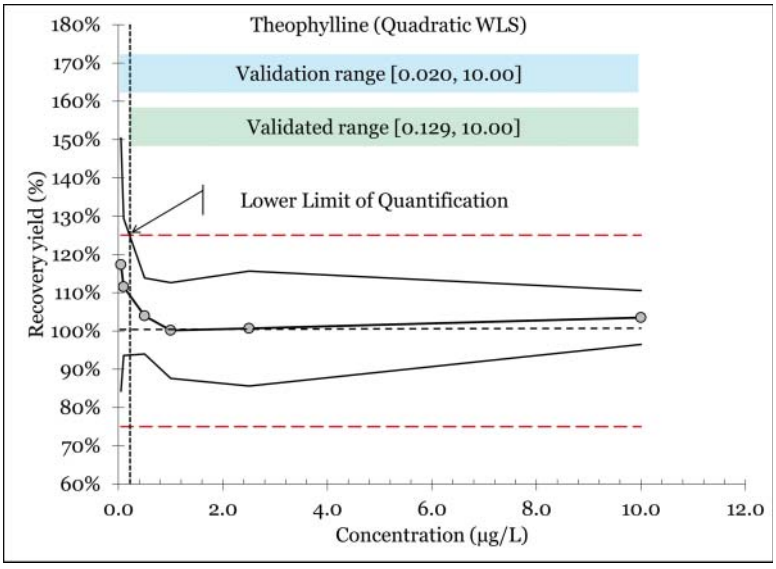



**Figure 5.6** THEOPHYLLINE – validation and validated ranges ($\beta$% = 80%).

Finally, the concentration where the method is no longer valid is determined by any intersection between one of the β-ETI interpolation segments and the acceptance interval limits. In this example, the crossing happens around 0.1–0.2 µg/l. More exactly, it is at 0.129 µg/l, and the rationale for accurately calculating this point is explained below. Two concentration ranges can then be defined:

- The validation range *a priori* set by the analyst, i.e. [0.020, 10.00] µg/l.
- The validated range is calculated by considering the intersection point between β-ETI and the acceptance interval, when it exists. For this example, the validated range spreads between [0.129, 10.00] µg/l. Some FDA regulatory documents also propose to define the two bounds of the validated range as the lower limit of quantification (LLOQ) and the upper limit of quantification (ULOQ), respectively.

An important point when calculating the intersection point to establish the LLOQ is to use the *absolute inverse-predicted concentrations*, as collected in Table 5.3, and no longer the recovery yields, as in Table 5.4. This precaution is compulsory to avoid a bias of the artifact due to the linear interpolation between TI bounds. Proper interpolation curves should be hyperbolas.

Figure 5.7 illustrates the lower part of the accuracy profile, between 0.02 and 0.6 µg/l. It shows how to calculate the intersection point. In mathematical terms, it consists in computing the intersection of two straight lines related to absolute inverse-predicted concentrations. This means finding the roots of a system of two equations with two unknowns. This is a typical algebraic problem easily solved with any worksheet. Obviously, when $\beta$ probability or acceptance interval changes, so is the intersection point modified as well as the bounds of the validated domain.

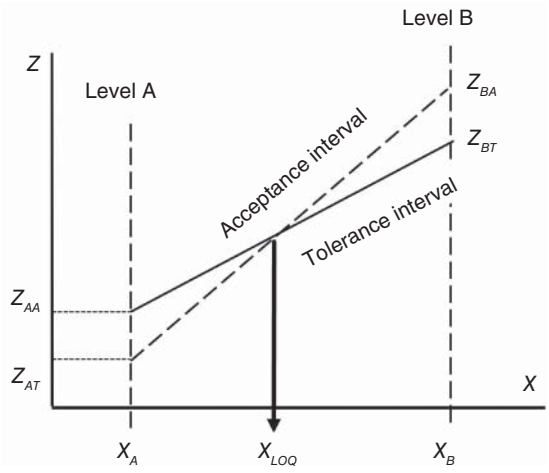



**Figure 5.7** THEOPHYLLINE – lower part of the MAP expressed as absolute inverse-predicted concentration ($\beta\% = 80\%$). The arrow indicates the lower limit of the validated range that can be used as *LOQ* (0.129 µg/l).

To illustrate the method, let us consider the data recorded in Table 5.3 for the THEOPHYLLINE example at levels $A = 0.1$ µg/l and $B = 0.5$ µg/l, since it is between these two levels that one segment of the tolerance interval profile crosses of the upper limit of acceptability. As stated, it is compulsory to use absolute values. The following notations reported in Figure 5.8 will help to understand the computation:

- $Z$ recovered concentration values on the $Y$-axis.
- $X$ known concentration values on the $X$-axis
- $X_A$ and $X_B$ concentrations of the selected levels $A$ and $B$, respectively.
- $Z_{AT}$ and $Z_{BT}$ tolerance interval upper bounds at selected levels.
- $Z_{AA}$ and $Z_{BA}$ acceptance interval upper bounds at selected levels.
- $X_{LOQ}$ concentration at the crossing point defined as the *LOQ*.

These two straight lines can be represented by a system of two equations, the first describes the acceptance limit, which varies with the concentration as it is defined as a percentage, and the second the tolerance interval interpolation segment. Since, by construction, the acceptance interval always passes through zero, we must find $A_0 = 0$.

| | |
|---|---|
| Equation (1.1) for the acceptance interval | $Z = A_0 + A_1 X$ |
| Equation (1.2) for the tolerance interval | $Z = T_0 + T_1 X$ |

The slope and intercept of Eq. (1.1) are calculated as follows:

| | |
|---|---|
| Slope of Eq. (1.1) | $A_1 = \frac{Z_{BA} - Z_{AA}}{X_B - X_A}$ |
| Intercept of Eq. (1.1) | $A_0 = Z_{AA} - A_1 X_A$ |
| Symmetrically for Eq. (1.2): | |
| Slope of Eq. (1.2) | $T_1 = \frac{Z_{BT} - Z_{AT}}{X_B - X_A}$ |
| Intercept of Eq. (1.2) | $T_0 = Z_{AT} - T_1 X_A$ |

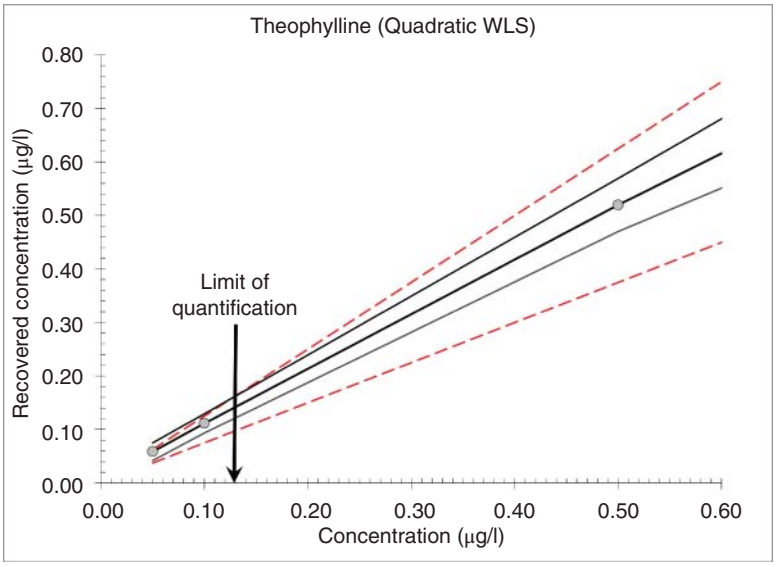



**Figure 5.8** Schematic representation for LOQ calculation.

Finally, the abscissa of the crossing point is given by:

$$X_{LOQ} = \frac{A_0 - T_0}{T_1 - A_1}$$

To illustrate the calculation, a worksheet can easily be implemented as illustrated. Finally, $X_{LOQ} = 0.129$ µg/l. It must also be noted that $A_0 = 0$ as expected. When the proportion $\beta$ is modified, so is the *LOQ*. For instance, when $\beta\% = 67\%$, it becomes $X_{LOQ} = 0.099$ µg/l. This remark confirms the usefulness of this proposal to define the *LOQ,* but harmonization is necessary to make it acceptable. Chapter 9 suggests a more extensive discussion about this fundamental validation parameter.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Limit of Quantification from the Accuray Profile** | | | |
| 2 | Levels | A | B | |
| 3 | Concentration (µg/L) | 0.100 | 0.500 | |
| 4 | Upper bounds Tolerance interval | 0.125 | 0.625 | |
| 5 | Upper acceptance limit | 0.129 | 0.569 | |
| 6 | Equation coefficients | | | |
| 7 | Tolerance interval | Slope | 1.250 | =(C4-B4)/(C3-B3) |
| 8 | | Intercept | 0.000 | =B4-C7*B3 |
| 9 | Acceptance interval | Slope | 1.099 | =(C5-B5)/(C3-B3) |
| 10 | | Intercept | 0.019 | =B5-C9*B3 |
| 11 | Limit Of Quantification | | 0.129 | =(C10-C8)/(C7-C9) |

## 5.3 Statistical Dispersion Intervals

An interval for $Z$ is a set of real numbers for which $a \leq Z \leq b$, where $a$ and $b$ are real numbers called the bounds or the limits of the interval. It is used together with the symbol $[a, b]$. Globally there are two categories of statistical intervals, one is used to describe, the other to predict.

For instance, confidence intervals (CI) are the best-known, widely explained and used in the statistical literature. It is a descriptive interval around the computed value of a parameter, such as a mean or a standard deviation. It is assumed to contain the theoretical *true* value of the corresponding parameter with a certain probability of error conventionally denoted $\alpha$. True value is likely to lie in this interval; the adjective "*likely*" relates to the confidence level of the confidence interval, classically denoted $1 - \alpha$. The practical definition of the CI is available in ISO 23833:2013: "range of analytical error expected to contain the true value with a stated uncertainty as estimated from a statistical model of the measurement process." Any true statistical parameter of a population is assumed to be bounded by the confidence interval.

In contrast, there are at least two other categories of statistical dispersion intervals that are less well known but extensively discussed here, namely the prediction intervals and the tolerance intervals (TI). In the prospect of method validation and measurement uncertainty, they are particularly interesting. Unfortunately, the definitions are confusing.

In statistical literature, there are references to both tolerance intervals, sometimes referred to as:

- β-γ-content tolerance interval (β-γ-CTI).
- β-expectation tolerance interval (β-ETI).

TIs were introduced by the end of the 1940s when quality assurance and statistical process control principles became widespread in industry. Table 5.5 proposes a classification of different statistical intervals noted from A to D, depending on the study's goal. It is compiled and adapted from [24].

The *prediction interval* is an interval that will "contain a future randomly selected observation from a distribution." With a specified degree of confidence. It is generally useful to predict the result of one, or a small number, of future measurements. Prediction intervals for all future observations are of interest only if a small number of measurements are produced because they are often very wide. Also, the exact number of future measurements is sometimes not known or may conceptually be infinite. Moreover, rather than requiring that the calculated interval contain a specified number of units, it is generally sufficient to construct an interval to contain a substantial proportion of such units.

**Table 5.5** Examples of some statistical intervals.

| Goal | Description | Prediction |
|---|---|---|
| Range | A. Tolerance interval to contain (or cover) at least a specified proportion of a distribution | B. Prediction interval to contain the observations from a future sample |
| Probability | C. Confidence interval for the probability of an observation being less than (or greater than) some specified value | D. Prediction interval to contain the proportion of observations in a future sample that exceed a specified limit |

Source: Adapted from Meeker et al. [24].

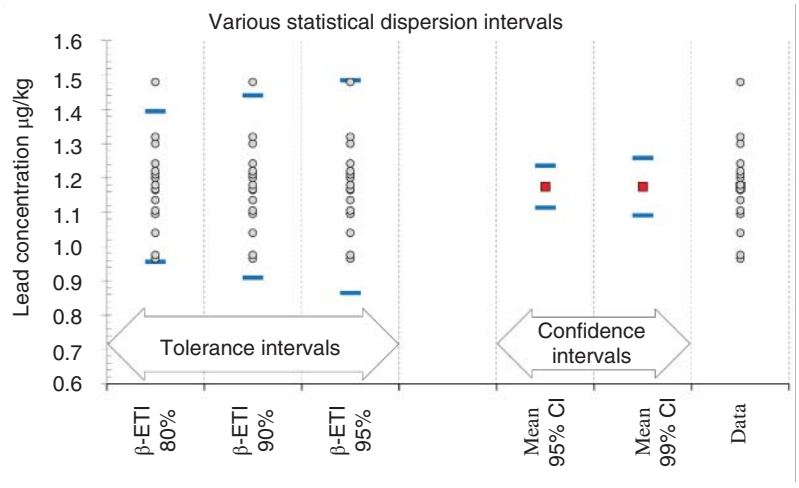



**Figure 5.9** Comparison tolerance and confidence intervals calculated for 18 replicates assumed to be normally distributed.

Specifically, the TI is an interval that one can claim to "contain at least a specified proportion, $\beta$, of the distribution" with a specified degree of confidence usually denoted $\gamma$ in this context but corresponding to the classic level of confidence used for CI. Such an interval is of particular interest in limiting the process capability for measurements produced in massive quantities. This contrasts with a prediction interval which is, as defined, of greatest interest in predicting a small number of future units.

*Thus, it was demonstrated that a β-expectation tolerance interval (β-ETI) can also be defined as a prediction interval for a single future observation.*

To illustrate how these distinct types of statistical intervals may apply, Figure 5.9 compares the intervals obtained with 18 replicates performed on a single sample and assumed to be normally distributed[1]. This is the simplest situation and formulas used for this example are not presented but can be found in many publications, such as [25]. The major difference between tolerance and confidence intervals appears: the TI is applied to describe the whole data distribution, while the CI is used to characterize a statistical parameter (the mean for this example) which is a combination of data.

Apparently, there is some confusion among the definitions of the term *tolerance* as used in standards or guidelines. In ISO 16269-6 standard, a TI is "statistical dispersion interval," and in ISO 12669-8 standard, it is a "prediction interval." Moreover, tolerance can also be defined as a complement to acceptance. For instance, in ISO/IEC Guide 98-4, the following definition is proposed for the tolerance interval "the interval of permissible values of a property" and, at the opposite, the rejection interval is "the interval of non-permissible measured values." In the same document, the acceptance interval is "the interval of permissible measured quantity values" [26]. The concepts are reviewed in Section 8.2 on the conformity assessment of a

1 Unpublished personal data

sample. In the following Sections 5.3.1 and 5.3.2, the concept of tolerance interval, abridged as TI, is used the same way as it is classically presented in the statistical literature. For the validation of analytical methods, two kinds of TI have been proposed:

| | |
|---|---|
| β-Expectation tolerance interval (β-ETI) | It contains the value of a future observation, with a coverage probability $\beta$. As already mentioned, this is the type of statistical interval chosen by the SFSTP commission for the calculation of a MAP and the validation of analytical methods (see Tables 5.3 and 5.4). It corresponds to the letter A in Table 5.5. As explained in Section 7.2, its great advantage is to be suitable for estimating MU since it corresponds to an interval containing a given proportion of the possible measurements of a measurand. This property fits the definition of the coverage interval introduced in Section 7.6. |
| β-γ Content tolerance interval (β-γ-CTI) | It is also known as the guaranteed coverage tolerance interval with a confidence level of $\gamma$ [27]. In this case, $\beta$ is the proportion as in β-ETI, while $\gamma$ is a confidence level. As specified in ISO 3534-1:2006, in this context, the confidence level is the "long-term proportion of intervals" constructed in this manner that will include at least the expected proportion. This type of statistical dispersion interval is more difficult to relate to the estimation of MU. On the other hand, it is adequately adapted to setting up a quality control or a control chart, as illustrated in Section 7.3. |

Tolerance interval estimators generally assume specific data distributions, such as Normal distribution exemplified by Figure 5.9, or Poisson distribution. The data collected to build a MAP has a more complex hierarchical structure due to the fact that measurements are completed in several series. The series is interpreted as a random effect factor, similar to the laboratory effect during an interlaboratory study (see Section 3.2). Therefore, the calculation of the TIs must take this structure into account. Each measurement value is considered as the combination of several distribution laws. The method for computing β-ETI described in the following chapters applies to this type of structured data and has been specifically developed by several authors for balanced and unbalanced experimental designs [27, 28].

### 5.3.1 β-Expectation Tolerance Interval (β-ETI)

Following our previous convention, $X$ represents the known concentration of a validation material, $Y$ the instrumental response, and $Z$ the inverse-predicted concentration, directly obtained or by inverting the calibration function. The β-ETI for a population of measurement values $Z$ is then expressed as follows:

$$[\overline{\overline{Z}} \pm k_{TI} \times s_{TI}]$$

with

$\overline{\overline{Z}}$: Grand mean or global average of all data ($I \times J$).

$k_{TI}$: Coverage factor of β-ETI depending on the coverage probability $\beta$%.

$s_{TI}$: (Combined) standard deviation of the β-ETI.

The following formulas are applicable for calculating the parameters of β-ETI:
Grand mean of inverse-predicted concentrations

$$\overline{\overline{Z}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} Z_{ij}}{I \times J} \tag{5.2}$$

Number of series

$$1 \le i \le I \tag{5.3}$$

Number of replicates per series

$$1 \le j \le J \tag{5.4}$$

Variance ratio

$$A = \frac{s_B^2}{s_r^2} \tag{5.5}$$

Weighting coefficient $Q$

$$Q = \frac{A+1}{J \times A + 1}$$

$$Q = IJ\left(\frac{\frac{s_B^2}{s_r^2}+1}{J \times \frac{s_B^2}{s_r^2}+1}\right) \tag{5.6}$$

Variance of intermediate precision (see note 1)

$$s_{IP}^2 = s_r^2 + s_B^2$$

Variance of the β-ETI

$$s_{TI}^2 = s_{IP}^2\left(1 + \frac{1}{I \times J \times Q}\right)^2 \tag{5.7}$$

Standard deviation of the β-ETI (see note 2)

$$S_{TI} = s_{IP}\sqrt{1 + \frac{1}{I \times J \times Q}} \tag{5.8}$$

Number of degrees of freedom ($df$) or Number of effective measures

$$N_E = \frac{(A+1)^2}{\frac{\left(A+\frac{1}{J}\right)^2}{I-1} + \frac{1-\frac{1}{J}}{IJ}} \tag{5.9}$$

Coverage factor (quantile of Student's law) $\quad k_{TI} = t_{N_E, \frac{1+\beta}{2}}$ (5.10)

Note 1: Formulas of variances $s_{IP}^2$, $s_r^2$ and $s_B^2$ are available in Section 3.4.1 and obtained with the classic standard ANOVA algorithm of ISO 5725.

Note 2: The standard deviation of the β-ETI is introduced here for convenience to simplify further computation but does not appear in the literature.

The parameter $N_E$ is considered as a number of *effective* measures because it can be used to check whether the total number of data $I \times J$ collected to build the interval is appropriate, as explained in Section 5.4.3. This number depends on many

intermediate parameters, mainly the variance ratio $A$. On the other hand, from a statistical viewpoint, it corresponds to the number of degrees of freedom after estimating the TI standard deviation. By definition, it can be interpreted as the number of independent values that remain after all the relationships that link data have been established. It plays a key role in computing the coverage factor $k_{TI}$.

Because the starting point is to compute the repeatability and intermediate precision variances, Resource E worksheet should be used to compute the β-ETI. The simplest solution is to add new lines and new formulas, leading to another resulting worksheet called Resource H. It is applied to the THEOPHYLLINE dataset on the subset of measurements obtained for the second validation material and listed in Table 5.2. In this example, the assigned reference value is identified as the known concentration of the validation material, that is 0.100 µg/l. For each validation material (or level), the same worksheet must be copied and updated with corresponding measurements.

The value in cell `B2` (yellow highlighted) containing the assigned value must be consequently modified, as well as the measurement values in cells `B5:D10`. Following the principle already explained, formulas are made visible in column `C` to allow the reader to adapt their own worksheet. They are on the right of column `B`, which contains the expected result. This worksheet is an example that can be improved or adapted. For instance, if there are more than six series of two replicates or if the design is unbalanced. Until row 30, all formulas are the same as in Resource E worksheet. The lines after row 42 will be explained in Section 7.2.4 about MU.

The clumsiest part of this worksheet is between rows 35 and 38, where the coverage factor is computed. Because the formula for obtaining the number of degrees of freedom ($df$) $N_E$ is complex, it is recommended to split it into several lines to better control possible typing errors.

After all, this number of $df$ is usually non-integer; for this example, it is 9.59. Equation (5.10) shows how the $df$ is used to calculate the quantile of the Student's $t$ distribution law which corresponds to the interval coverage factor. Unfortunately, the Excel built-in function `TINV` which gives the quantile of the Student's $t$, only accepts integer numbers of $df$. However, it is essential to know this quantile to obtain the coverage factor.

In the absence of a suitable function, an approximate value should be obtained by linear interpolation between the quantiles given by the upper and lower-rounded integers of $N_E$. In Section 5.4.3, Figure 5.15 illustrates the differences between exact and approximate Student's $t$ value obtained by linear interpolation. Obviously, this approximation may generate an error that is significant if $N_E < 4$.

We will see that the minimal requirement for building a MAP is to collect at least $I \times J = 9$ measures by level, and it is exceedingly rare to cope with this downside. In such situations, it is recommended to use the `t.ppf` Python function available after importing the `scipy.stats` package that gives exact values for non-integer numbers of $df$. Lines 42–45 include formulas to calculate the MU. Details about this new quantity are available in Section 7.2, where all equations are explained and commented.

**Resource H β-Expectation Tolerance Interval (Excel).**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Resource H: β-Expectation Tolerance Interval** | | | | | |
| 2 | Assigned Reference Value (µg/L) | 0.100 | | | | |
| 3 | | Recovered concentration | | | | |
| 4 | | Replicate 1 | Replicate 2 | n(i) | SS(i) | |
| 5 | Series 1 | 0.114 | 0.113 | 2 | 5.00E-07 | |
| 6 | Series 2 | 0.112 | 0.110 | 2 | 2.00E-06 | |
| 7 | Series 3 | 0.104 | 0.101 | 2 | 4.50E-06 | |
| 8 | Series 4 | 0.100 | 0.105 | 2 | 1.25E-05 | |
| 9 | Series 5 | 0.113 | 0.107 | 2 | 1.80E-05 | |
| 10 | Series 6 | 0.147 | 0.112 | 2 | 6.13E-04 | |
| 11 | **General parameters** | | | | | |
| 12 | Number of series (I) | 6 | =COUNT(B5:B10) | | | |
| 13 | Number of measures (IJ) | 12 | =COUNT(B5:C10) | | | |
| 14 | Number of replicates (J) | 2 | =IF(D5*B12<>B13;"Error";D5) | | | |
| 15 | Residual sum of squares (SSW) | 0.00065000 | =SUM(E5:E10) | | | |
| 16 | Total sum of squares (SSt) | 0.00163500 | =DEVSQ(B5:C10) | | | |
| 17 | Inter-series sum of squares (SSB) | 0.00098500 | =B16-B15 | | | |
| 18 | Repeatability variance ($s^2r$) | 0.00010833 | =B15/(B13-B12) | | | |
| 19 | Temporary between-variance | 0.00004433 | =((B17/(B12-1))-B18)/B14 | | | |
| 20 | Between-series variance ($s^2B$) | 0.00004433 | =IF(B19<0;0;B19) | | | |
| 21 | Reproducibility variance ($s^2R$) | 0.00015267 | =B18+B20 | | | |
| 22 | **Precision** | | | | | |
| 23 | Recovered concentration | 0.112 | =AVERAGE(B5:C10) | | | |
| 24 | Repeatability std. dev. (sr) | 0.0104083 | =SQRT(B18) | | | |
| 25 | Between-series std. dev. (sB) | 0.0066583 | =SQRT(B20) | | | |
| 26 | Intermediate Precision std. dev.. (sFI) | 0.0123558 | =SQRT(B21) | | | |
| 27 | **Trueness** | | | | | |
| 28 | Relative bias (%) | 11.5% | =(B23/B2)-1 | | | |
| 29 | Recovery yield (%) | 111.5% | =B23/B2 | | | |
| 30 | **beta-Expectation Tolerance Interval** | | | | | |
| 31 | Tolerance (beta) | 80% | | | | |
| 32 | Variance Ratio (A) | 0.409 | =B20/B18 | | | |
| 33 | Coefficient Q | 0.7750 | =(B32+1)/(B14*B32+1) | | | |
| 34 | Weighing factor W | 1.0524 | =SQRT(1+1/(B13*B33)) | | | |
| 35 | Number of degrees of freedom (NE) | 9.59 | =(B32+1)^2/((B32+1/B14)^2/(B12-1)+(1-1/B14)/B13) | | | |
| 36 | t Student lower | 1.38 | =TINV(1-B31;ROUNDDOWN(B35;0)) | | | |
| 37 | t Student upper | 1.37 | =TINV(1-B31;ROUNDUP(B35;0)) | | | |
| 38 | Coverage factor (kTI) | 1.38 | =B36-(B36-B37)*(B35-ROUNDDOWN(B35;0)) | | | |
| 39 | b-ETI standard deviation (sTI) | 0.013003 | =B26*B34 | | | |
| 40 | Lower bond expectation interval | 0.094 | =B23-B38*B39 | | | |
| 41 | Upper bond expectation interval | 0.129 | =B23+B38*B39 | | | |
| 42 | **Uncertainty** | | | | | |
| 43 | Composed standard uncertainty (uc(Z)) | 0.013 | =B39 | | | |
| 44 | Extended uncertainty (U(Z)) | 0.026 | =2*B43 | | | |
| 45 | Relative uncertainty (UR%) | 26.0% | =B44/B2 | | | |

### 5.3.2 β-γ Content Tolerance Interval (β-γ-CTI)

Another type of statistical dispersion interval is called β-γ-CTI and is considered by several authors to be well-suited to validate analytical methods [27]. The benefit of β-γ-CTI is to allow an easy computation of the warning and control limits of a control chart, as explained in Section 4.4, but the downside is that it is not suitable for estimating MU. The formulas for the β-γ-CTI are combined below, using the same notations as β-ETI.

**Table 5.6** THEOPHYLLINE – parameters of β-γ-CTI ($\beta\% = 80\%$, $\gamma\% = 95\%$).

| Reference value (µg/l) | | 0.05 | 0.10 | 0.50 | 1.00 | 2.50 | 10.00 |
|---|---|---|---|---|---|---|---|
| Std. dev. of the interval | $s_{IC}$ | 0.0214 | 0.0227 | 0.0642 | 0.1634 | 0.4555 | 0.8940 |
| Weighting coefficient | $W$ | 7.22 | 9.30 | 7.23 | 6.41 | 12.93 | 8.91 |
| Coverage factor | $k_{IC}$ | 1.37 | 1.35 | 1.37 | 1.38 | 1.33 | 1.35 |
| Content interval bounds | | 0.03 | 0.08 | 0.43 | 0.78 | 1.91 | 9.14 |
| | | 0.09 | 0.14 | 0.61 | 1.23 | 3.12 | 11.56 |

General formula $\quad \overline{\overline{Z}} \pm k_{IC} \times s_{IC}$ (5.11)

Coverage factor of the interval $\quad k_{IC} = z_{\left(\frac{1+\beta}{2}\right)} \times \sqrt{1 + \frac{1}{W}}$ (5.12)

Quantile of Normal law $\quad z_{\left(\frac{1+\beta}{2}\right)}$ (5.13)

Standard deviation of the interval

$$s_{IC} = \sqrt{s_{FI}^2 + \sqrt{\left(\frac{H_1 \times s_B}{J}\right)^2 + \left(\frac{H_2 \times (J-1) \times s_r}{J}\right)^2}} \tag{5.14}$$

Weighting coefficient $\quad W = \dfrac{I \times \left\{\left(\frac{SCE_B}{IJ-1}\right) + (J-1)\left(\frac{SCE_r}{I-1}\right)\right\}}{\left(\frac{SCE_B}{IJ-1}\right)}$ (5.15)

Coefficient $H_1 \quad H_1 = \dfrac{I-1}{\chi^2_{1-\gamma,I-1}} - 1$ (5.16)

Coefficient $H_2 \quad H_2 = \dfrac{I(J-1)}{\chi^2_{1-\gamma,I(J-1)}} - 1$ (5.17)

The notation $\chi^2_{1-\gamma,I(J-1)}$ represents the quantile of the $\chi 2$ distribution (chi-square) for the probability $1-\gamma$ and the number of degrees of freedom equal to $I(J-1)$. This value is directly obtained with the Excel built-in function `CHIINV`. Table 5.6 groups the parameters obtained from the THEOPHYLLINE data provided in Table 5.2. Likewise, the β-ETI, before plotting these new intervals on the accuracy profile, the values relative to the reference concentration value $X$ must be calculated. In Table 5.7 table, the relative limits of the β-γ-CTI as well as those of β-ETI and the acceptance limits, are collected for comparison.

The complete MAP for theophylline is illustrated in Figure 5.10. The bounds of the β-γ-CTI are larger as they bring in a confidence level. They can be interpreted as intervals where the *true* β-ETI may probably lie with a risk of error of 5%. The worksheet about precision parameters referenced Resource E was extended to give a new worksheet applicable to the computation of the β-ETI and called Resource H. Likewise, the worksheet called Resource I β-γ Content Tolerance Interval (Excel) for the β-γ-CTI is an extension of Resource H and uses the same layout.

**Table 5.7** THEOPHYLLINE – summary of relative tolerance intervals bounds.

| | | | | | | |
|---|---|---|---|---|---|---|
| Reference value (µg/l) | 0.05 | 0.10 | 0.50 | 1.00 | 2.50 | 10.00 |
| Recovery (%) | 117% | 112% | 104% | 100% | 101% | 104% |
| β-ETI $\beta\% = 80\%$. | 84% | 94% | 94% | 88% | 86% | 96% |
| | 150% | 129% | 114% | 113% | 116% | 111% |
| β-γ-CTI with $\beta\% = 80\%$; $\gamma\% = 95\%$. | 59% | 81% | 86% | 78% | 76% | 91% |
| | 176% | 142% | 121% | 123% | 125% | 116% |
| Acceptance limit ± 25% | 75% | 75% | 75% | 75% | 75% | 75% |
| | 125% | 125% | 125% | 125% | 125% | 125% |

**Resource I** β-γ **Content tolerance interval (Excel).**

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | **Resource I: β-γ-Content Tolerance Interval** | | | | |
| 2 | Assigned Reference Value (µg/L) | 0.100 | | | |
| 3 | | Recovered concentration | | | |
| 4 | | Replicate 1 | Replicate 2 | ni | SSi |
| 5 | Series 1 | 0.114 | 0.113 | 2 | 5.00E-07 |
| 6 | Series 2 | 0.112 | 0.110 | 2 | 2.00E-06 |
| 7 | Series 3 | 0.104 | 0.101 | 2 | 4.50E-06 |
| 8 | Series 4 | 0.100 | 0.105 | 2 | 1.25E-05 |
| 9 | Series 5 | 0.113 | 0.107 | 2 | 1.80E-05 |
| 10 | Series 6 | 0.147 | 0.112 | 2 | 6.13E-04 |
| 11 | **General parameters** | | | | |
| 12 | Number of series (I) | 6 | =COUNT(B5:B10) | | |
| 13 | Number of measures (IJ) | 12 | =COUNT(B5:C10) | | |
| 14 | Number of replicates (J) | 2 | =IF(D5*B12<>B13;"Error";D5) | | |
| 15 | Residual sum of squares (SSW) | 0.0006500 | =SUM(E5:E10) | | |
| 16 | Total sum of squares (SSt) | 0.0016350 | =DEVSQ(B5:C10) | | |
| 17 | Inter-series sum of squares (SSB) | 0.0009850 | =B16-B15 | | |
| 18 | Repeatability variance ($s^2r$) | 0.0001083 | =B15/(B13-B12) | | |
| 19 | Temporary Between variance | 0.0000443 | =((B17/(B12-1))-B18)/B14 | | |
| 20 | Between-series variance ($s^2B$) | 0.0000443 | =IF(B19<0;0;B19) | | |
| 21 | Reproducibility variance ($s^2R$) | 0.0001527 | =B20+B18 | | |
| 22 | **Precision** | | | | |
| 23 | Recovered concentration | 0.1115 | =AVERAGE(B5:C10) | | |
| 24 | Repeatability std. dev. (sr) | 0.01041 | =SQRT(B18) | | |
| 25 | Between-series std. dev. (sB) | 0.00666 | =SQRT(B20) | | |
| 26 | Intermediate Precision std. dev.. (sFI) | 0.01236 | =SQRT(B21) | | |
| 27 | **Trueness** | | | | |
| 28 | Relative bias (%) | 11.5% | =(B23/B2)-1 | | |
| 29 | Recovery yield (%) | 111.5% | =B23/B2 | | |
| 30 | **beta-gamma-Content Tolerance Interval** | | | | |
| 31 | Tolerance (beta) | 80% | | | |
| 32 | Confidence (gamma) | 95% | | | |
| 33 | Repeatability mean square | 0.00011 | =B15/(B12*(B14-1)) | | |
| 34 | Between-series mean square | 0.00020 | =B17/(B12-1) | | |
| 35 | Weighting coefficient (W) | 9.29949 | =(B12*(B33+(B14-1)*B34))/B34 | | |
| 36 | Normal law quantile | 1.282 | =NORMSINV((1+B31)/2) | | |
| 37 | Coverage factor (kIC) | 1.349 | =B36*SQRT(1+(1/B35)) | | |
| 38 | Coefficient H1 | 3.365 | =((B12-1)/CHIINV(B32;B12-1))-1 | | |
| 39 | Coefficient H2 | 2.669 | =((B12*(B14-1))/CHIINV(B32;B12*(B14-1)))-1 | | |
| 40 | Weighing coefficient (W) | 0.00036 | =SQRT(((B34*B38/B14)^2)+((B33*B39*((B14-1)/B14))^2)) | | |
| 41 | CTI standard deviation (sIC) | 0.02268 | =SQRT((B33+MAX(0;(B34-B33)/B14)+B40)) | | |
| 42 | Lower bound content interval | 0.081 | =B23-B37*B41 | | |
| 43 | Upper bound content interval | 0.142 | =B23+B37*B41 | | |

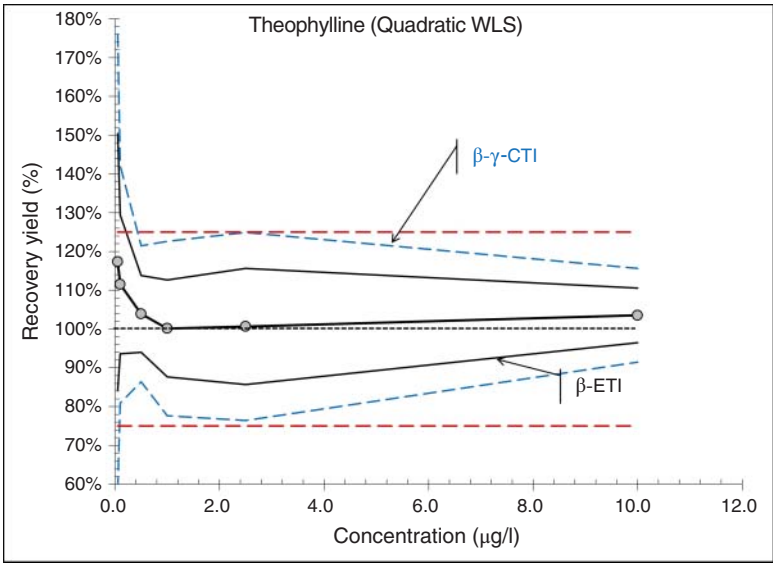



**Figure 5.10** THEOPHYLLINE – accuracy profile with the two tolerance intervals ($\beta\% = 80\%$, $\gamma\% = 95\%$). Inverse-predicted concentrations are obtained with WLS quadratic models.

In Resource I worksheet, the calculation of the standard deviation of the interval, denoted $s_{IC}$. is a little bit complicated. It requires two weighting coefficients, $H_1$ and $H_{2,}$ which contain the quantiles of the $\chi^2$ law that are calculated as follows:

| Coefficient | Degrees of freedom | Excel formulas | Value |
|---|---|---|---|
| $H_1$ | $I - 1 = 6 - 1 = 5$ | `=CHIINV(0.95;5)` | 3.36 |
| $H_2$ | $I \times (J - 1) = 6 \times 1 = 6$ | `=CHIINV(0.95;6)` | 2.67 |

Up to row 30, all formulas are the same for both TIs. Therefore, it is possible to combine them into one single worksheet. In rows 31 and 32, the two probability values associated with this type of interval are entered. Some formulas are quite long and made visible in column C. The worksheet could be improved by reorganizing Eqs. (5.14)–(5.17). It is also possible to put together in a single worksheet the three programs Resource E, Resource H, and Resource I. To present the details of the calculation and facilitate the adaptation and verification by any laboratory, they are kept separate. In Section 5.4.4, the influence of the choice of probabilities $\beta$ and $\gamma$ is discussed in more detail, especially in view of the various official validation guides.

## 5.4 Accuracy Profile: Special Topics

Besides its application for method validation, MAP can also be of some help in optimizing some final steps of the operating procedure or checking if the planning of the validation study was correct or may be improved.

**Table 5.8** THEOPHYLLINE – coefficients of all quadratic models adjusted to the series of calibration data.

| Series | Algorithm | $a_0$ | $a_2$ | $a_3$ | $r^2$ (%) | AIC |
|---|---|---|---|---|---|---|
| 1 | OLS | 1.0403 | 11.779 | 0.0837 | 99.5 | 59.50 |
| 1 | WLS | 0.0489 | 16.312 | −0.4281 | 95.5 | 39.46 |
| 2 | OLS | 0.5482 | 15.258 | −0.1924 | 100.0 | 23.87 |
| 2 | WLS | 0.0375 | 17.828 | −0.485 | 99.4 | 20.94 |
| 3 | OLS | 0.1614 | 17.506 | −0.4985 | 99.9 | 39.04 |
| 3 | WLS | 0.1866 | 17.697 | −0.5232 | 99.0 | 25.05 |
| 4 | OLS | 0.3039 | 6.5998 | −0.0847 | 99.9 | 30.69 |
| 4 | WLS | 0.1408 | 7.2525 | −0.1574 | 99.4 | 3.67 |
| 5 | OLS | 0.1874 | 8.5221 | −0.118 | 100.0 | −1.09 |
| 5 | WLS | 0.0391 | 9.2555 | −0.2014 | 99.7 | 0.86 |
| 6 | OLS | 0.3315 | 6.4809 | −0.0855 | 100.0 | 11.69 |
| 6 | WLS | 0.0398 | 7.7705 | −0.2306 | 99.3 | 5.56 |

### 5.4.1 Choose the Best Calibration Model

In Section 2.1, it was mentioned that various calibration models can be adjusted to the same dataset, using either OLS or WLS estimators. The first question was to select the best calibration model. In statistical literature, different parameters are proposed, such as the coefficient of determination $r^2$ or the Akaike information coefficient (AIC). Regarding the example in Table 2.2, no decision could be made because these parameters are too close. In other words, they are not sensitive enough. Table 5.8 summarizes the coefficients for all quadratic models separately adjusted to the six-calibration series of the THEOPHYLLINE dataset. These values were obtained using the short Python script of Resource B applied to the five other series of calibration data (not provided).

The situation seems even more confusing as the calibration model coefficients are highly variable. It is even surprising to have consistent inverse predicted concentrations when the coefficients seem to change so much. This confirms two basic proposals:

- Validation study should be organized into several series of measurements to verify that the quantification capacity of the method is not affected over time by various sources of variation.
- Calibration model is a critical issue but not the key issue. Using calibration parameters, such as sensitivity or linearity, is not pertinent to assess method validated range.

It is possible to construct several MAPs with different inverse-predicted concentrations obtained either by OLS or WLS calibration models. When differences

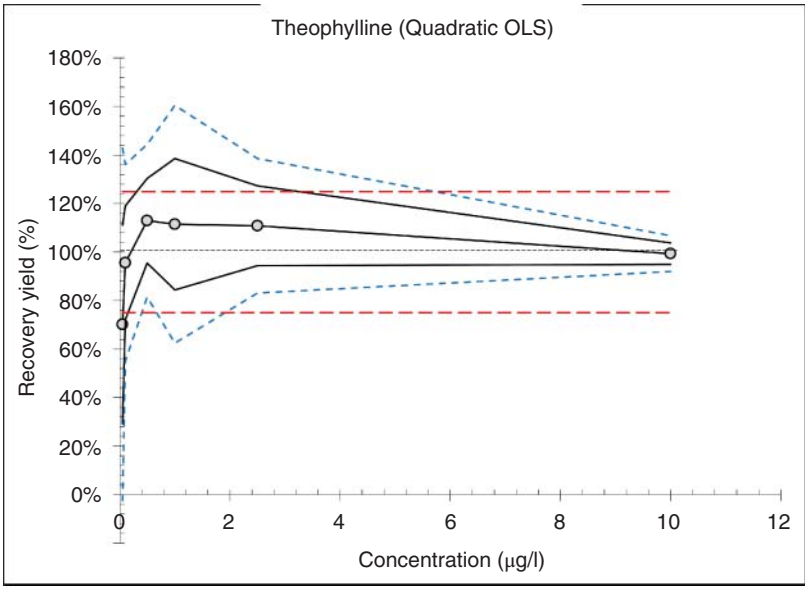



**Figure 5.11** THEOPHYLLINE – accuracy profile (MAP) with the two tolerance intervals obtained for OLS quadratic models ($\beta\% = 80\%$, $\gamma\% = 95\%$).

appear among these profiles, it may help to make the final decision about the best calibration model. The profile based on OLS models is illustrated in Figure 5.11 and must be compared to Figure 5.5 obtained with WLS models. The decision is evident: the new MAP is much less favorable than the previous one as the validated range is clearly diminished; the OLS model is unsatisfactory. Moreover, a proposal was to define the *LOQ* as the lowest limit of the validated range. It was 0.129 µg/l with WLS curves, while it is at 2.698 µg/l with the new MAP. This example illustrates the possible use of MAP as an ultimate optimization tool. For future routine application of the theophylline method of analysis, for each series, the calibration curve *must* be a second-order polynomial calculated by weighted least-squares
regression.

In Section 2.3.2 about WLS, it was underlined that the choice of the weighting may have some consequences on the performance of the inverse prediction. All presented computations were done using the questionable $1/X^2$ weighting factor. Even better results may be obtained with another more adapted choice.

### 5.4.2 Apply Consistent Experimental Design

Since the 1930s, many published books focused on the theory of experimental design. The early contribution of R.A. Fisher (1890–1962), called *The design of experiments* is the keystone of the theory. In the introduction, it is explained why, in any experimental work, it is fundamental to organize the trials according to well-established rules:

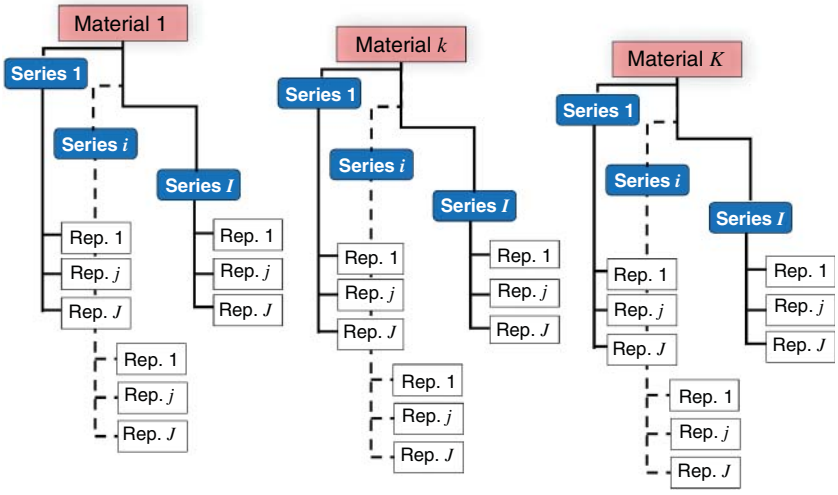



**Figure 5.12** Schematic representation of the experimental design to be used to build a relevant accuracy profile. Material means reference material of known concentration.

> …a criticism frequently levelled at test results is that the experiment was poorly designed and, therefore, poorly conducted. Assuming that the experimenter did what he intended to do, this point comes down to the question of the design and logical structure of the experiment.

Since this work, there has been much evidence that the relevance of experimental data fundamentally depends on the conditions it was collected. A consequent number of publications propose optimal experimental designs adapted to different contexts [29]. Therefore, to obtain a relevant accuracy profile, it is logical also to use a relevant experimental design based on perquisite rules. Figure 5.12 exemplifies these rules.

Four rules are proposed to design a consistent validation experiment.

---

Rule 1. Select at least three validation materials ($K \geq 3$) with known (or known with a defined uncertainty) concentration levels covering the validation range. They can combine different matrices as far as they have comparable effects. The number of three is minima as it is used to verify the trueness is linear, i.e. the proportionality exists between the known contents of reference material and the recovered contents after application of the whole analytical procedure. A greater number of validation materials would be preferable, as only three may lead to disappointment.

Rule 2. Perform under intermediate precision conditions at least three series of measurements ($I \geq 3$) for each validation material, i.e. so that several sources of uncertainty, as representative as possible of the routine condition, come into play from one series to another.

Rule 3. Provide a minimum of two replicates per series ($J \geq 2$) on the same validation material, performed under repeatability conditions. A number of official guides have a higher requirement ($J = 3$ to 6).

Rule 4 (optional but allows for simplified calculations later). Ensure all series have the same number of replicates to have a balanced experimental design.

---

For example these rules align with the experimental design requirements for drug development studies advocated in ISO or FDA standards, but not the ICH Q2(R2) guidelines [3] that recommend a minimum number of replicates $J = 3$. The THEOPHYLLINE study is a good example of such a balanced design that can be summarized in a condensed form as follows:

| | |
|---|---|
| $I = 6$ | Number of measurement series for each material, performed over six days. |
| $J = 2$ | Number of replicates per series. |
| $K = 6$ | Number of validation materials with known target concentration levels, set between 0.20 and 10.0 µg/l. |

We propose the concise notation $6s/2r/6l$ for this design, where $s$ stands for the series, $r$ for replicates, and $l$ for levels. In this example, the choice results from a compromise between the elution time for each run and the need to establish the MAP over a very wide concentration range.

### 5.4.3 Check the Number of Efficient Measurements

When the experimental design is in harmony with the above-stated rules, it is possible to compute the number of effective measurements denoted $N_E$ with formula (5.9) presented above and recalled here.

$$N_E = \frac{(A+1)^2}{\frac{\left(A+\frac{1}{J}\right)^2}{I-1} + \frac{1-\frac{1}{J}}{IJ}} \quad \text{With} \quad A = \frac{s_B^2}{s_r^2}$$

$N_E$, corresponds to the estimated number of $df$, according to the well-established Satterthwaite's approximation procedure [30]. The probability $\beta$ and this number of $df$ are used to obtain the quantile of the Student's $t$ distribution law which appears in the formula of the coverage factor of the β-ETI (Eq. 5.10). Finally, $N_E$ depends on three parameters: two are chosen *a priori* by the experimenter, i.e. $I$ is the number of series, and $J$ is the number of replicates per series; the third is the variance ratio $A$ of the between-series variance to the repeatability variance (or within-series variance) and is only known *a posteriori* at the end of the experiment.

Being the number of $df$, $N_E$ is the number of independent quantities which can be assigned to a parameter estimate and be statistically interpreted as the number of measures that efficiently contribute to the β-ETI. The higher this number, the more reliable the estimate and, in this case, the smaller the coverage factor. It is, therefore, an important indicator of what can be called to as the "quality of the study." Most importantly, in terms of scientific or financial optimality, $N_E$ allows us to check whether the parameters of the experimental design have been correctly chosen and whether the explanation of the β-ETI intervals is satisfactory.

In addition, $N_E$ also lets us complete the experimental design at the best cost by calculating the optimal number of runs that would eventually have to be added to obtain narrower intervals. In view of its role, it is interesting to simulate the values that $N_E$ takes as a function of the three parameters on which it depends, $I$, $J$, and $A$.

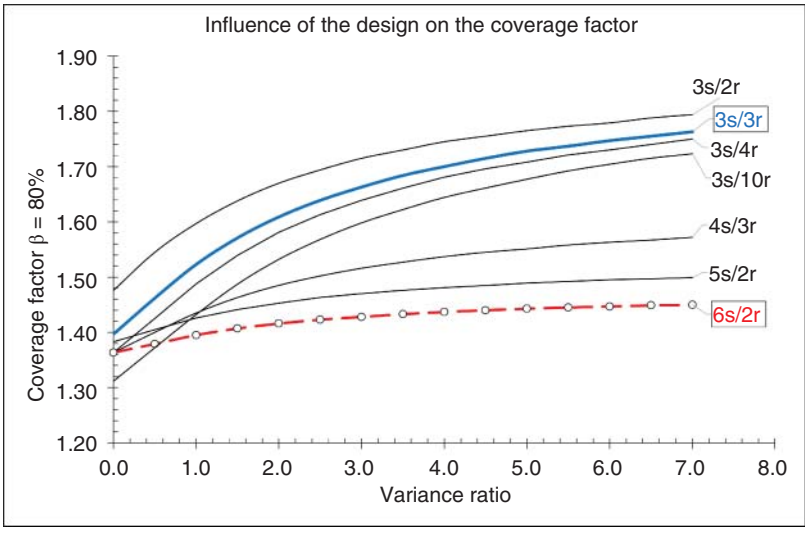



**Figure 5.13** Influence of the experimental design parameters on the number of effective measurements. A design is identifiable by *s* the number of series and *r* the number of replicates per series, i.e. *Is*/*Jr*.

Figure 5.13 illustrates these variations as a function of the variance ratio for different combinations of $I$ and $J$ marked by $s$, the number of series and $r$, the number of replicates, in the form $Is/Jr$.

It is obvious that when the variance ratio $A \to 0$, the number of efficient measurements tends towards $(IJ-1)$, while $A$ plays a fundamental role and underlines significant differences between the number of trials and the number of values finally carrying information, whatever the various combinations of $I$ and $J$.

In several official guides, the classically recommended minimum configuration is labeled $3r/3s$. In Figure 5.13, it appears as a thick solid line. As soon as the variance ratio $A > 3$, $N_E$ falls very quickly from 8 to 3, meaning that almost two-thirds of the data do not participate in the estimation of TIs. It is remarkable that the addition of 1 replicate per series ($3s/4r$) to this design does not drastically improve the situation. In conclusion, when the between-series variance is high compared to the repeatability, this "official" design is unfavorable. It is possible to define the most effective design for a given budget.

For instance, if it is decided to pay for 12 analyses per level, the optimal configuration is $6s/2r$, or 6 series of 2 replicates/series, since $N_E$ always remains higher than 5. In reverse the $2s/6r$ combination appears as the least interesting since $N_E$ falls below 2 when $A$ is increasing. In other words, less than 20% (2/12) of data are informative. The abacus of all other 12-trial designs, listed from bottom to top, $2s/6r$, $4s/3r$, $3s/4r$, and $6s/2r$, confirms this trend.

The conclusion is clear: for a fixed budget, the most often successful strategy is to make as many series as possible at the expense of the number of replicates. It is also possible to check the limits of variation of the coefficient $Q$, whose calculation

**Table 5.9** Asymptotic values of the number of efficient measures for special values of the variance ratio *A*.

| Variance ratio | $A \to 0$ | $A = 1$ | $A \to +\infty$ |
|---|---|---|---|
| Number of efficient measures $N_E$ | $N_E \to (IJ-1)$ | | $N_E \to I-1$ |
| Coefficient $Q$ | $Q \to 1$ | $Q = \frac{2}{J+1}$ | $Q \to \frac{1}{J}$ |

**Table 5.10** THEOPHYLLINE – efficiency rates of the experimental design.

| Concentration (µg/l) | | 0.05 | 0.1 | 0.5 | 1 | 2.5 | 10 |
|---|---|---|---|---|---|---|---|
| Number of measures | $IJ$ | 12 | 12 | 12 | 12 | 12 | 12 |
| Optimal number of *df* | | 11 | 11 | 11 | 11 | 11 | 11 |
| Observed number of *df* | $N_E$ | 7.01 | 9.59 | 7.02 | 5.69 | 10.91 | 9.22 |
| Efficiency rate | $R_E$ | 64% | 87% | 64% | 52% | 99% | 84% |
| Variance Ratio | $A$ | 1.95 | 0.41 | 1.93 | 6.78 | 0.00 | 0.53 |

is provided by the formula (5.18).

$$\text{Coefficient } Q \quad Q = \frac{A+1}{J \times A + 1} \tag{5.18}$$

When $A$ varies from 0 to $+\infty$, Table 5.9 gathers the extrema of $N_E$ and $Q$ as a function of some special values of $A$. Recall that if $A \to 0$, the between-series variance $s_B^2$ also tends to 0, $s_B^2 \to 0$. It means that there is no between-series effect, the intermediate precision standard deviation is equal to the repeatability standard deviation, and the conditions of quantification remain constant from one series to the next; this is the most favorable situation. The opposite conclusion is reached when $A \to +\infty$.

It is then possible to propose a method to verify whether the observed dataset will result in a satisfactory accuracy profile, by calculating the efficiency rate $R_E$ of the experiment. It is obtained by ratioing the optimal number of efficient measurements, i.e. $(IJ-1)$ when $A \to 0$ and the number of efficient measurements $N_E$ estimated from the data. It can be expressed as percent:

$$\text{Efficiency rate} \quad R_E = \frac{N_E}{IJ-1} \times 100 \tag{5.19}$$

Table 5.10 summarizes the various parameters calculated from the data of the THEOPHYLLINE study and extracted from Table 5.3. It confirms former remarks:

- At level 2.5 µg/l, $A = 0$ and $N_E = 10.91$ close to $(IJ-1) = 11$, study efficiency is almost 100%.
- At level 1.0 µg/l, $A = 6.78$, and $N_E = 5.69$ is close to $I-1 = 5$ and the efficiency falls to 52%.

Unfortunately, the variance ratio $A$ is only known at the end of the experiment, whereas its influence is important. To overcome this drawback, the organization of

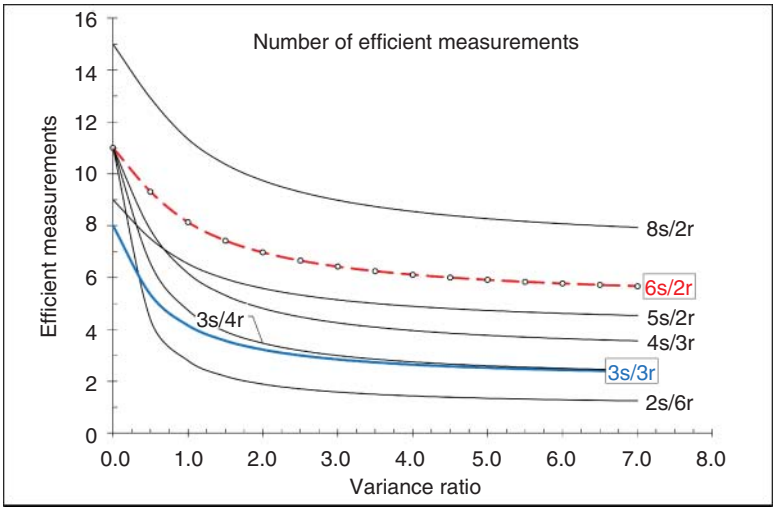



**Figure 5.14** Influence of the experimental design on the coverage factor with $\beta\% = 80\%$.

the experimental design can be considered as an evolutionary process. Initially, a favorable configuration is chosen, for example, $5s/2r$, and the trials are conducted. Then, the number of series can eventually be increased according to the values of the ratio $A$ and by taking, as an optimality criterion, the coverage factor of the interval, which is given by the following formula:

$$\text{Coverage factor (quantile of the Student's law)} \quad k_{IT} = t_{N_E, \frac{1+\beta}{2}} \tag{5.20}$$

This parameter depends on $N_E$, and consequently, on the $A$ ratio. Using observed data, it is possible to predict how many additional series it would be interesting to add to the previous ones to reduce the coverage factor as much as possible and remain within the budget. The only constraint is to keep $J$ the number of replicates/series constant so that the design remains balanced.

Figure 5.14 uses the same design labeling convention as Figure 5.13 to illustrate the consequences of the coverage factor reduction when the number of series is increased rather than the number of replicates. Once the design contains six series, the variance ratio has almost no influence. Once the variance ratio $A$ is estimated, it is possible to simulate the consequences of adding new measurement series to the existing ones and check whether the coverage is reduced.

Such simulations using the parameters obtained at the issue of the THEOPHYLLINE study are compiled in Table 5.11. Initially, the design included six series; two were added, and the new coverage factors were computed, supposing that the variance ratio remained constant. The reduction of the β-ETI width is small, about 2–4%, and it is not interesting to add more measurements. It can be explained by the small values of the ratio $A$ and the good experiment efficiency rates $R_E$, which confirms that this validation is satisfactory.

The publication on theophylline also presents the results obtained with the same method applied to caffeine quantification [23]. The experimental design $6s/2r$ is

**Table 5.11** Simulation of the addition of 2 series of measurements to the initial experimental design.

| **Theophylline** | | | | | | |
|---|---|---|---|---|---|---|
| Concentration (µg/l) | 0.05 | 0.1 | 0.5 | 1 | 2.5 | 10 |
| Number of series (I) | 6 | 6 | 6 | 6 | 6 | 6 |
| Number of replicates (J) | 2 | 2 | 2 | 2 | 2 | 2 |
| Variance ratio | 1.95 | 0.41 | 1.93 | 6.78 | 0.00 | 0.53 |
| Initial coverage factor 80% | 1.41 | 1.38 | 1.41 | 1.45 | 1.36 | 1.38 |
| Added series | 2 | 2 | 2 | 2 | 2 | 2 |
| Updated $N_E$ | 9.79 | 13.30 | 9.81 | 7.96 | 14.93 | 12.81 |
| Updated coverage factor 80% | 1.37 | 1.35 | 1.37 | 1.40 | 1.34 | 1.35 |
| Decrease in β-ETI width | −3% | −2% | −3% | −4% | −2% | −2% |
| **Caffeine** | | | | | | |
| Concentration (µg/l) | 0.02 | 0.15 | 0.5 | 1 | 2.5 | 10 |
| Variance Ratio | 25.20 | 1.14 | 1.17 | 6.81 | 0.79 | 0.00 |
| Initial $N_E$ | 5.20 | 7.90 | 7.80 | 5.70 | 8.60 | 10.90 |
| Initial coverage factors 80% | 2.55 | 2.31 | 2.32 | 2.49 | 2.28 | 2.20 |
| Updated $N_E$ | 7.3 | 11.0 | 11.0 | 8.0 | 11.9 | 14.9 |
| Updated coverage factor 80% | 1.41 | 1.36 | 1.36 | 1.40 | 1.36 | 1.34 |
| Decrease in β-ETI | −45% | −41% | −41% | −44% | −40% | −39% |

the same, but only validation material reference values vary. For theophylline, the addition of two new series is not especially useful, while for caffeine, it is much more profitable because variance ratios are much higher. In this case, the β-ETI could is reduced up to 45% at a low cost, as illustrated in the second part of Table 5.11.

The variation in the coverage factor $k_{IT}$ as a function of the number of efficient measurements, i.e. the number of $df$ is illustrated in Figure 5.15 for three classic values of the $\beta$. It corresponds to the quantile of the Student's $t$ distribution law for noninteger numbers of $df$. As explained in Section 5.3.1 about Resource H, it can be obtained by two methods: an approximate value using the built-in Excel function called `TINV`; or an exact value using Python `t.ppf` function.

Both results are reported in Figure 5.15, with the exact value as a continuous line and the interpolate as a broken dashed line. When the number of $df$ is below 4, the differences between both values can be significant. To summarize, some recommendations can be extracted from this figure to check if the validation study went well, but they must be balanced according to the application being treated:

- Is the number of effective measures $N_E \geq 5$ ?
- Is the variance ratio $A \leq 2$ ?

The variance ratio $A$ can also be interpreted as a robustness parameter over time as it measures the between-series variability over the total variability. The performance of the method can be assumed as stable if $A$ is small when shifting from one series to

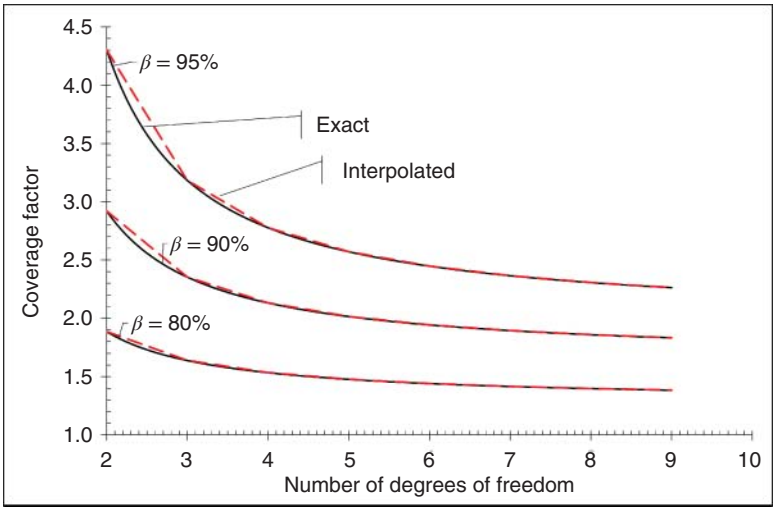



**Figure 5.15** Coverage factor as a function of the number of efficient measurements.

another or over time if series are scattered on different days. That is, the quantification procedure is not sensitive to slight variations between series. In the example of caffeine reported in Table 5.11, the $A$ ratio varies from one validation material (level) to another, while it is more stable for theophylline. The analytical technique used, namely UHPLC-MS-MS, is known to be delicate to use, which probably explains these variations.

### 5.4.4 Select Probability Values

Whether β-ETI or β-γ-CTI is used, three values must be selected to construct and interpret the accuracy profile:

- $\beta$: The coverage probability, i.e. the content proportion of future measurements of the tolerance interval, may be bound.
- $\gamma$: The confidence level for the β-γ-CTI, which represents the probability of a confidence interval. In statistical literature, the confidence level is usually denoted $1-\alpha$ where $\alpha$ is the risk of error. When using $\gamma$ instead, we are just conforming to the classic literature on statistical dispersion intervals.
- $[A_L; A_U]$: The acceptance interval, also expressed as a percentage and defined by a probability of success.

The acceptance interval is centered on the nominal value of the validation material but is not always equal on both sides as some regulations may require asymmetric values, such as a recovery yield ranging between 80% and 110%, as detailed in Table 5.16. Regarding tolerance intervals, they are centered on the average inverse-predicted concentration.

This means they are generally symmetrical. But for some analytical methods, where the results are previously transformed into logarithms, the antilog-intervals

are no longer symmetrical. This situation is observable for polymerase chain reaction (PCR) methods applied to the detection of genetically modified organisms (GMO).

The choice of the three values listed above is application domain-specific and may not be the same for the methods applicable to the environment, consumer health, pharmaceutical industry, etc. For example, consider the values published by the FDA in the guidance on *Bioanalytical Method Validation* [6]. Two categories of analytical methods frequently employed in bioanalysis are considered: chromatographic assays (CC) and ligand binding assays (LBAs). The requirements of the section called "In study analysis" may help to define the three probability values in this case.

The FDA requirements are expressible in the compact form 4/6/15% for chromatographic assays and 4/6/20% for immuno-analyses. They are decoded as follows: four out of six quality controls (QC), or 67%, must be within ±15% (or ±20%) acceptance interval around the nominal contents of the material used for the quality control. In other words, it is acceptable that a proportion of 67% of measurements made on a QC lie in the interval ±15% (or ±20%) around the nominal value. β-γ-CTI can also be used to establish some acceptance rules for the analysis quality control required by the US-FDA in the context of biological analysis.

Summary of Table 1 of FDA Guidance *Bioanalytical Method Validation*.

| |
|---|
| *Quality Controls (QC)* |
| Elements: |
| – ≥4 QC levels (LLOQ, low, medium, and high) and ≥2 replicates per QC level in each analytical run. |
| – Total QCs should be 5% of unknown samples or ≥6, whichever number is greater. |
| – If the analytical runs consist of distinct processing batches, the QC acceptance criteria should be applied for the whole run and each distinct batch within the runs. |
| Acceptance criteria: |
| – CC: ≥67% (4/6) of QCs should be ±15% of the nominal, and ≥50% of QCs per level should be ±15% of their nominal. |
| – LBA: ≥67% (4/6) of QCs should be ±20% of the nominal, and ≥50% of QCs per level should be ±20% of their nominal. |
| **Trueness**[a] **and precision** |
| Trueness: between runs: |
| – CC: ±15% of nominal concentrations |
| – LBA: ±20% of nominal concentrations |
| Precision: between runs: |
| – CC: ±15% CV |
| – LBA: ±20% CV |

a) Beware: in the original document the term "accuracy" is used instead of "trueness," conversely to the VIM definitions.

Source: Adapted from FDA [6].

**Table 5.12** THEOPHYLLINE – half tolerance intervals for different probability values.

| | Concentration (µg/l) | | | | | |
|---|---|---|---|---|---|---|
| **Parameters** | **0.05** | **0.1** | **0.5** | **1** | **2.5** | **10** |
| Acceptance probability | 20% | 20% | 20% | 20% | 20% | 20% |
| ITE-β (80%) | 33.1% | 17.9% | 9.9% | 12.5% | 15.0% | 7.0% |
| ITC-β-γ (80–95%) | 58.6% | 30.6% | 17.6% | 22.5% | 24.2% | 12.1% |
| ITC-β-γ (67–95%) | 44.6% | 23.2% | 13.3% | 17.1% | 18.4% | 9.2% |

The names of the intervals are abbreviated as in the text and the associated probabilities appear in brackets.

Official FDA requirement is that at least 4 out of 6, i.e. 67% of measurements made on QC samples, fall within the acceptance interval $[A_L; A_U]$, which is, depending on the type of method, either ±15% or ±20% of the reference (or nominal) value.

To have a margin of safety and to be sure of achieving this objective, it may be recommended to choose a probability of 80% per level so that at least 4 out of 5 or 80% (instead of more than 4 out of 6) measurements are included in the chosen acceptance interval. Finally, according to FDA rules, the probability values associated with the statistical dispersion intervals, which are used to verify whether a validated method is suitable for QC, come down to the following:

- Content proportion $\beta$%: 80% or 67%.
- Confidence level $\gamma$%: 95%.
- Acceptance: ±15% or ±20%

Table 5.12 combines the half-intervals, expressed in % and calculated from the THEOPHYLLINE dataset, for various combinations of probability and interval types. To restore the coverage of one of the intervals in measurement units, simply apply the percentage to the concentration in the top row. For example, the half β-ETI (80%) is 33.1% for a concentration of 0.05 µg/l, which gives the following bonds for the coverage interval for 80% of future measurements:

$$[0.05 \times (1 - 0.331); 0.05 \times (1 + 0.331)] = [0.033; 0.067]$$

Figure 5.16 shows this set of values in graphical form, which shows the respective roles of the coverage probability $\beta$ and the confidence level $\gamma$. If the three concentration levels of the QCs are correctly chosen, then this method of determining theophylline perfectly satisfies the FDA requirements for quality control. Indeed, between 0.13 and 10 µg/l, the half intervals are below the acceptance limit symbolized by the dashed line at ±20%. From this example, it is also clear that changing $\beta$% from 67% to 80% has a significative influence on the width of the interval at all concentrations.

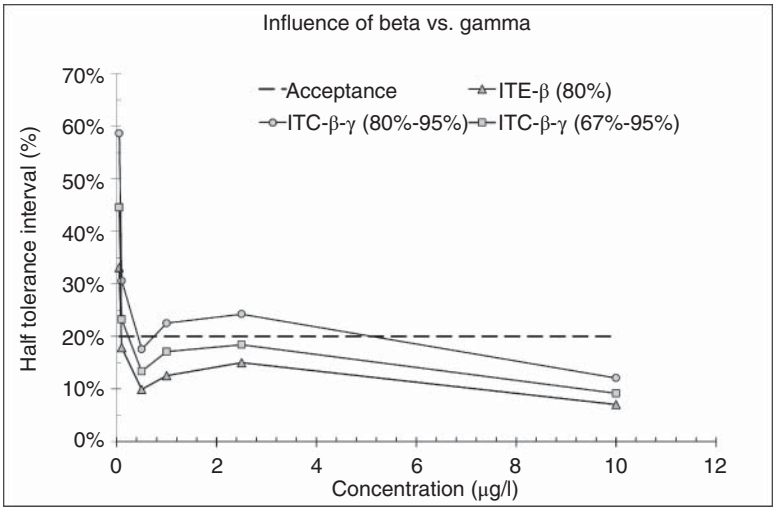



**Figure 5.16** THEOPHYLLINE – half tolerance intervals for different probability values. Probabilities are in brackets after interval name abbreviations.

### 5.4.5 Select the Type of Tolerance Interval

Referring to the literature, both types of intervals-β-ETI and β-γ-CTI-have been proposed to conduct method validation. The following table summarizes the main arguments and literature now attached to either type of interval. In the last row appears the type of TI recommended in this book.

| | **β-Expectation interval (β-ETI)** | **β-γ-Content interval (β-γ-CTI)** |
|---|---|---|
| Objective | Predict the interval that contains a known proportion of future observations. | Define the limits of a specified proportion of a distribution with a guarantee. |
| References | [17–20] | [27] |
| Applications | Method validation and/or estimation of MU | Control charts and/or routine quality control |

As previously mentioned, the β-ETI is thus particularly suited to predicting an interval computed from a set of observations that contains a given proportion of future measurements. It therefore, predicts whether the method can produce, for a given sample, a high proportion of acceptable measurements, i.e. within the limits of a predefined acceptance interval. In practice, a recommended value of β% = 80% was adopted because it allows compliance with the requirement formulated by the FDA, as explained in the previous chapter. In the Section 7.3, which deals with control charts, it is also shown that this proportion allows us to follow the typical requirements encountered for standardized control charts.

### 5.4.6 Proportion of Nonacceptable Measures

A classic misunderstanding about the $\beta$ proportion is to believe that only $\beta$% acceptable measurements values are produced by the method. This is an incorrect interpretation of this probability. The correct interpretation is that, at least, $\beta$% results are within the tolerance interval if the β-ETI remains within the acceptance interval bounds. While this condition is true, much more than $\beta$% produced measurement values are acceptable.

To illustrate this point, it is possible to predict the proportion of acceptable measurement values that will be inside the acceptance interval $[A_L, A_U]$. Consider the parameters listed in Table 5.3 page 118 for the THEOPHYLLINE study. The second validation material contains 0.5 µg/l, and the acceptance interval is ±25%, giving:

| | |
|---|---|
| Inverse-predicted grand mean | $\overline{\overline{Z}} = 0.520$ |
| Limits of the acceptance interval | $[A_L, A_U] = [0.5 - 25\% ; 0.5 + 25\%]$ |
| | $[A_L, A_U] = [0.375, 0.625]$ |
| Standard deviation of β-ETI | $s_{IT} = 0.0350$ |
| Number of degrees of freedom | $N_E = 7.02$ |

The method consists in calculating the differences between the inverse-predicted grand mean and the two limits of the acceptance interval. If they are divided by the standard deviation of the β-ETI, these two quantities denoted $t_{inf}$ and $t_{sup}$ can be considered as approximate quantiles of a Student's law with $N_E$ degrees of freedom. In this example, $N_E = 7.02$, and allows to obtain the probability associated with each quantile.

| Quantile of the Student's *t* law | Associated probability |
|---|---|
| $t_{sup} = \frac{\lvert\overline{\overline{Z}} - A_U\rvert}{s_{IT}} = \frac{0.1458}{0.0350} = 4.1292$ | 0.22% |
| $t_{inf} = \frac{\lvert A_L - \overline{\overline{Z}}\rvert}{s_{IT}} = \frac{0.1054}{0.0350} = 3.0106$ | 0.98% |

When the theophylline concentration is around 0.5 µg/l, the global risk of producing non-acceptable measures is $0.98\,\% + 0.22\,\% = 1.2\%$ which is much below the traditional value of 5% used for statistical testing. It also means that, at this concentration, the method can produce about 98.8% of acceptable measures.

Table 5.13 generalizes this calculation for the different concentration levels. The global risk of *nonacceptable* results very rapidly decreases while concentration increases. The breakpoint of more than 25% of nonacceptable measures lies between 0.05 and 0.1 µg/l where the LOQ is located. To achieve this calculation, Student's law table that provides the probability associated with noninteger number of degrees of freedom is necessary.

As explained for Resource H worksheet it is possible to obtain an approximate value by interpolation using the built-in function `TDIST` that gives this probability. `TINV` is the inverse function of `TDIST`. To obtain the percentage of non-acceptable

**Table 5.13** THEOPHYLLINE: - probabilities of non-acceptable measurement values (Acceptance interval ± 25%).

| | Concentration (µg/l) | | | | | |
|---|---|---|---|---|---|---|
| **Parameters** | **0.05** | **0.10** | **0.50** | **1.00** | **2.50** | **10.00** |
| Lower quantity | 1.8110 | 2.8070 | 4.1292 | 2.9164 | 2.3335 | 5.6003 |
| Higher quantity | 0.3280 | 1.0382 | 3.0106 | 2.8854 | 2.2141 | 4.2173 |
| Interval std. dev. | 0.0117 | 0.0130 | 0.0350 | 0.0862 | 0.2749 | 0.5093 |
| Number of *df* | 7.01 | 9.59 | 7.02 | 5.69 | 10.91 | 9.22 |
| Below lower bounds | 5.65% | 0.97% | 0.22% | 1.44% | 1.99% | 0.02% |
| Above upper bounds | 37.63% | 16.23% | 0.98% | 1.49% | 2.45% | 0.11% |
| Global risk | 43.28% | 17.20% | 1.20% | 2.93% | 4.45% | 0.12% |

measurements, the simplest method consists in adding the following lines and formulas to the end of the Resource H worksheet as shown in Resource J. It is applied to the theophylline level of 0.1 µg/l as the rest of the program. 0.97% of measurements that are below the limit of 0.075 µg/l, i.e. the lower limit of acceptance, and 16.23% exceed 0.125 µg/l. These results can be interpreted as an assessment of the risk taken by the analyst to produce results that would not meet the requirements of the end-user, bearing in mind that this is a probability and not a certainty!

**Resource J Probability of nonacceptable measurements (Excel).**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 47 | **Resource J: Probability of non acceptable results** | | | | | |
| 48 | Assigned Reference Value (µg/L) | 0.100 | =B2 | | | |
| 49 | Recovered concentration | 0.1115 | =B23 | | | |
| 50 | b-ETI standard deviation (sTI) | 0.013003 | =B39 | | | |
| 51 | Number of degrees of freedom (NE) | 9.59 | =B35 | | | |
| 52 | Acceptance proportion | 25% | | | | |
| 53 | Acceptance lower bound | 0.075 | =B48*(1-B52) | | | |
| 54 | Acceptance upper bound | 0.125 | =B48*(1+B52) | | | |
| 55 | Lower quantile | 2.807 | =(B49-B53)/B50 | | | |
| 56 | Upper quantile | 1.038 | =(B54-B49)/B50 | | | |
| 57 | Probability Student L | 1.0% | =TDIST(B$55;ROUNDDOWN($B$51;0);1) | | | |
| 58 | Probability Student U | 0.9% | =TDIST(B$55;ROUNDUP($B$51;0);1) | | | |
| 59 | Percentage of too low values | 0.97% | =B$57-((B$57-B$58)*($B$51-ROUNDDOWN($B$51;0))) | | | |
| 60 | Probability Student L | 16.3% | =TDIST(B$56;ROUNDDOWN($B$51;0);1) | | | |
| 61 | Probability Student U | 16.2% | =TDIST(B$56;ROUNDUP($B$51;0);1) | | | |
| 62 | Percentage of too high values | 16.23% | =B$60-((B$60-B$61)*($B$51-ROUNDDOWN($B$51;0))) | | | |

To conclude, the question of the choice of the type of tolerance interval.

– To conduct the validation of a method, we recommend using the β-ETI because it allows to highlight a set of parameters, such as the number of effective measurements or the variance ratio, which are especially useful at optimizing

the experimental design. A method for estimating MU can also be easily derived from this type of TI by modifying the probability $\beta\%$ as explained in Section 7.2.

– The advantage of the β-γ-CTI is that it includes a confidence level, which is useful for knowing the bounds of the entire population of measurements for the same sample. This is the situation encountered with control charts where the same reference material is going to be analyzed several times in a row in the context of quality control. With a coverage probability of $\beta\% = 80\%$ and a confidence level $\gamma\% = 95\%$, it is possible to define, for a given concentration level, the bounds of an interval where, for 95% of 4 QC out of 5 would lie. According to FDA recommendations, it is even acceptable to take $\beta\% = 67\%$ and $\gamma\% = 95\%$ so that 2 QC out of 3 will fall within the acceptance interval [6]. But the algebraic form of β-γ-CTI does not allow to define a strategy for optimizing the experimental design.

## References


1 2002/657/EC. (2002). Commission Decision of 12 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results (Text with EEA relevance) (notified under document number C(2002) 3044) (OJ L 221 17.08.2002, p. 8). http://data.europa.eu/eli/dec/2002/657/oj (accessed 1 September 2023).

2 Health and Consumer Protection Directorate (DG-SANCO) (2021). Document SANCO No. 11312/2021. *Analytical Quality Control and Method Validation Procedures for Pesticide Residues Analysis in Food and Feed.*

3 ICH (2022). International Council for Harmonisation of technical requirements for pharmaceuticals for human use (ICH) Guideline Q2(R2) on validation of analytical proceduresStep 2b. Amsterdam.

4 United States Pharmacopeia (USP) (2009). Chapter <1225> Validation of Compendial Procedures. http://www.uspbpep.com/usp29/v29240/usp29nf24s0_c1225.htmL (accessed 1 September 2023).

5 Association of Official Analytical Chemists (AOAC) (2016). *AOAC Official Methods of Analysis, Annex F: Guidelines for Standard Method Performance Requirements.* Arlington, USA.

6 Food and Drug Administration (FDA) (2018). *Bioanalytical Method Validation Guidance for Industry*. Washington, DC: Office of Communications, Division of Drug Information Center for Drug Evaluation and Research https://www.fda.gov/files/drugs/published/Bioanalytical-Method-Validation-Guidance-for-Industry.pdf (accessed 1 September 2023).

7 Chandran, S. and Singh, R.S.P. (2007). Comparison of various international guidelines for analytical method validation. *Pharmazie* 62: 4–14.

8 Raposo, F. and Ibelli-Bianco, C. (2020). Performance parameters for analytical method validation: controversies and discrepancies among numerous guidelines. *Trends in Analytical Chemistry* 129: 115913.

9 Bureau International des Poids et Mesures (BIPM) (2012). International vocabulary of metrology — Basic and general concepts and associated terms (VIM3).

JCGM 200:2012, BIPM, Sèvres, France. https://www.bipm.org/ (accessed 23 July 2023).

**10** (2017). Codex Alimentarius: Guidelines on performance criteria for methods of analysis for the determination of pesticide residues in food and feed (2017) CXG 90-2017. Several application documents are issued. For more specific information, consult https://www.fao.org/fao-who-codexalimentarius/codex-texts/guidelines/en/.

**11** Thompson, M., Ellison, S.L.R., and Wood, R. (2002). Harmonized guidelines for single-laboratory validation of methods of analysis. *Pure Applied Chemistry* 74 (5): 835–855.

**12** ICH (2022). International Council for Harmonisation of technical requirements for pharmaceuticals for human use (ICH) Guideline Q2(R2) on validation of analytical procedures, Step 2b. Amsterdam. https://www.ich.org/page/quality-guidelines (accessed 1 September 2023).

**13** Rozet, E., Ceccato, A., Hubert, C. et al. (2007). Analysis of recent pharmaceutical regulatory documents on analytical method validation. *Journal of Chromatography A* 1158: 111–125.

**14** Miller, J.N., Miller, J.C., and Miller, R.D. (2018). *Statistics and Chemometrics for Analytical Chemistry*, 6e. England: Pearson Education Limited.

**15** Standard ISO 17025:2017. *General requirements for the competence of testing and calibration laboratories*. ISO, Genève.

**16** Association of Official Analytical Chemists (AOAC) (2016). *AOAC Official Methods of Analysis, Annex F: Guidelines for Standard Method Performance Requirements*. Arlington, USA. https://academic.oup.com/aoac-publications/book/45491/chapter/392387882 (accessed 1 September 2023).

**17** Hubert, P., Nguyen-Huu, J.J., Boulanger, B. et al. (2004). Harmonization of strategies for the validation of quantitative analytical procedures: a SFSTP proposal – Part I. *Journal of Pharmaceutical and Biomedical Analysis* 36 (3): 579–586.

**18** Hubert, P., Nguyen-Huu, J.J., Boulanger, B. et al. (2007). Harmonization of strategies for the validation of quantitative analytical procedures: a SFSTP proposal – Part II. *Journal of Pharmaceutical and Biomedical Analysis* 45 (1): 70–81.

**19** Hubert, P., Nguyen-Huu, J.J., Boulanger, B. et al. (2007). Harmonization of strategies for the validation of quantitative analytical procedures: a SFSTP proposal – Part III. *Journal of Pharmaceutical and Biomedical Analysis* 45 (1): 82–96.

**20** Hubert, P., Nguyen-Huu, J.J., Boulanger, B. et al. (2008). Harmonization of strategies for the validation of quantitative analytical procedures: a SFSTP proposal – Part IV. Example of application. *Journal of Pharmaceutical and Biomedical Analysis* 48 (3): 760–771.

**21** Westgard, J.O. and Westgard, S.A. (2013). Total Analytic Error from Concept to Application. *Clinical Laboratory News*. https://www.aacc.org/publications/cln/articles/2013/september /total-analytic-error.

**22** Bureau International des Poids et Mesures (BIPM) (2012). *International vocabulary of metrology — Basic and general concepts and associated terms* (VIM3). JCGM 200, Clause 2.13. BIPM. https://www.bipm.org/ (accessed 23 July 2023).

**23** Gassner, A.I., Schappler, J., Feinberg, M., and Rudaz, S. (2014). Derivation of uncertainty functions from validation studies in biological fluids: application to the analysis of caffeine and its major metabolites in human plasma samples. *Journal of Chromatography A* 1353: 121–130.

**24** Meeker, W.Q., Hahn, G.J., and Escobar, L.A. (2017). *Statistical Intervals: A Guide for Practitioners and Researchers*, 2e. Wiley.

**25** NIST/SEMATECH (2023). *e-Handbook of Statistical Methods*. http://www.itl.nist.gov/div898/handbook/ (accessed 1 September 2023).

**26** ISO/IEC Guide 98-4:2012. *Uncertainty of measurement — Part 4: Role of measurement uncertainty in conformity assessment*. ISO, Genève.

**27** Hoffman, D. and Kringle, R. (2005). Two-sided tolerance intervals for balanced and unbalanced random effects models. *Journal of Biopharmaceutical Statistics* 15 (2): 283–293.

**28** Mee, R.W. (1984). β-Expectation and β-content tolerance limits for balanced one-way ANOVA random model. *Technometrics* 26 (3): 251–254.

**29** Box, G.E.P., Stuart-Hunter, J., and W.G. (2005). *Hunter: Statistics for Experimenters: Design, Innovation, and Discovery*, 2e. Wiley.

**30** Sahai, H. and Ojeda, M.M. (2004). *Analysis of Variance for Random Models: Theory, Methods, Applications, and Data Analysis*. Birkhäuser Boston Inc.

# 6

# Measurement Uncertainty (MU)

## 6.1 Principle of Measurement Uncertainty

Metrology is "the science of measurement." Its modern history goes back to 1875, with the signature of the Metre Convention and the creation of the Bureau International des Poids et Mesures (BIPM). The Metre Convention is an international treaty, currently signed by about 64 countries which establishes the BIPM as the world authority for action in the field of metrology. The practical role of BIPM is to develop and control international measurement standards. They are required to cover an ever-expanding range of applications and to provide proof of equivalence between the national standards of different countries.

BIPM is also in charge of developing and managing the International System (SI) of units. Unfortunately for analysts, it was not until 1971 that, during the 14th BIPM meeting, it was finally decided to address the chemical measurement issue and introduce a new unit, the mole, about 100 years after other existing units. This delay illustrates the current challenges of chemical metrology.

In 1977, several scientists pointed out in a note that there was no consensus among measurers on the expression of the *error of measurement*. It was then asked by the BIPM to address this problem with other national metrology institutes and to issue a recommendation. A detailed questionnaire was prepared covering the topics involved and circulated to 32 national metrology institutes recognized as having an interest in the subject, supplemented by five international organizations.

The importance of having an internationally accepted procedure for expressing a new concept called measurement uncertainty (MU) was recognized by all participants. It then remained to construct the appropriate method to achieve this objective. A working group was created dedicated to this purpose and it prepared a recommendation that led to the *Guide to the expression of uncertainty in measurement*, abbreviated GUM, published in 1995. The emergence of MU is thus the result of a conceptual evolution that spans almost two decades. The current version of the GUM dates from 2020 but is currently under revision [1].

Before MU was considered relevant to analytical sciences, the concept of total analytical error (*TAE*) had been introduced around 1974 [2]. The reasoning behind this parameter is rather classical, as it is a perfect illustration of what is called the "error approach" in GUM.

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

To define TAE, it is assumed that the deviation of the measurement from the true value of the sample – called total error – consists of two parts, one systematic and the other random. In practical terms, the bias is considered as an estimate of the systematic error, and the standard deviation of precision combined with a coverage factor is an estimate of the random error dispersion. A more recent and practical definition of *TAE* given by International Council for Harmonization (ICH) is "the sum of the absolute value of the errors in accuracy (%) and precision (%)," where accuracy must be identified as trueness according to International Vocabulary of Metrology (VIM) definition [3].

From a mathematical point of view, the *TAE* is summarized by Eq. (6.1):

Total analytical error

$$TAE = \delta + k_{TAE} \times s_{TAE} \tag{6.1}$$

- $\delta$ the estimate of the lack of trueness in the form of a bias, which is supposed to be constant and systematic.
- $s_{TAE}$ the estimate of precision (classically the intermediate precision standard deviation) assumed to cover the random part of measurement error.
- $k_{TAE}$ the coverage factor accounting for the probability of coverage.

This approach has been considered attractive for its simplicity in clinical biology, as evidenced by guidelines in this domain of application [4]. But this attractiveness is more intellectual than practical, and *TAE* has not only advantages. The separate assessment of random and systematic errors is generally required by regulatory bodies and raises many practical difficulties. If the separation were possible, these two quantities would not have the same statistical property with respect to uncertainty.

If it seems understandable to regard the standard deviation of precision as related to a random variable (or a combination of random variables). It is more difficult to consider the bias of trueness as constant only because it is qualified as systematic. In practice, the bias, as it can be estimated in the analytical sciences, is not constant and varies from one replicate to another. These inconsistencies may explain the relatively limited use of *TAE* in other fields of analysis than clinical biology. It also explains why BIPM (and measurement's specialists) abandoned the error approach, in favor of the "uncertainty approach" as explained in Section 4.1.2.

## 6.2 General Procedure to Estimating MU

The principle of the GUM, a general procedure for estimating MU, is relatively simple, as we shall show [1]. It is called the GUM general procedure in the following pages. Several organizations have published documents, through application examples, to adapt them to the analytical sciences and the needs of laboratories [5]. Unfortunately, it became quickly evident that this adaptation is neither always flexible nor practical.

The general GUM procedure for estimating the MU is shown in Figure 6.1. It consists of four steps that will serve as a roadmap for the various examples of calculation explained in the following chapters. It is useful to introduce some notations.

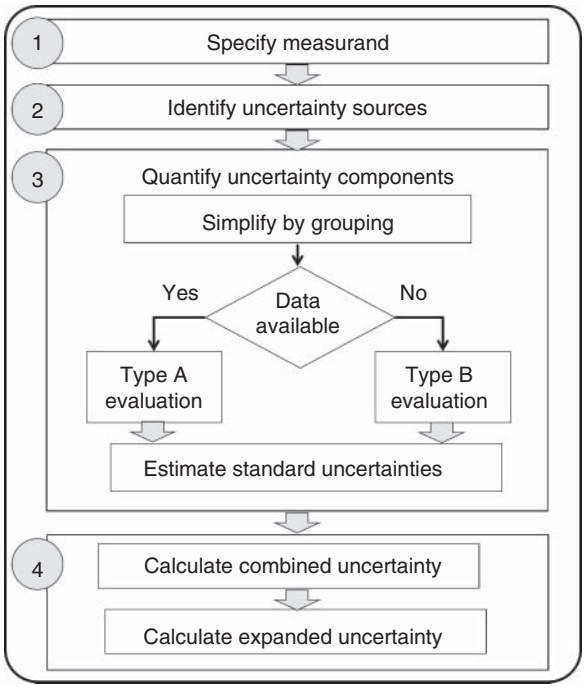



**Figure 6.1** The 4-step GUM general procedure for measurement uncertainty (MU) estimation.

Let us keep $Z$ to denote, as usual, the measured quantity of the analyte. MU can be expressed as four compatible parameters. Following the GUM conventions the following notations are used where $k_{GUM}$ is the coverage factor:
Composed standard uncertainty

$$u_c(Z) \tag{6.2}$$

Expanded uncertainty

$$U(Z) = k_{GUM} \times u_c(Z) \tag{6.3}$$

Relative uncertainty

$$UR\%(Z) = \frac{U(Z)}{Z} \times 100 \tag{6.4}$$

Coverage interval $[I_L; I_U]$

$$Z \pm k_{GUM} \times u_c(Z)$$
$$Z \pm U(Z) \tag{6.5}$$

The traditional values taken by the coverage factor $k_{GUM}$ are defined in the GUM and presented below. Because analysts are inclined to use percentages in expressing MU (and many other parameters), the relative uncertainty is part of the list but not always used in other measuring domains. The $L$ and $U$ subscripts of the coverage interval are abbreviations of lower and upper, respectively.

## 6.3 Traceability at the International System of Units

When estimating the MU of a result, an important preliminary issue is the attachment of the measurand to the International System of Units (SI), i.e. how the traceability of measurement values to official standards is ensured. For chemistry and biology, as stated, the SI unit of measure is the mole, whose symbol has been mol since 1970. Earlier, in the middle of the nineteenth century, the molar mass of the isotope 12 of carbon, denoted $M(^{12}C)$ was fixed at exactly 12.00000 g/mol. From there, by applying different methods of measurement, it was acknowledged by convention that a mole contained about $6 \times 10^{23}$ elementary entities. This number was called the Avogadro number at the beginning of the twentieth century. It is noted $N_A$ and initially came with a limited number of significant figures and an uncertainty that regularly decreased as measurement techniques improved. Until 2018, this reference to $^{12}C$ was the accepted definition of the mole. But at the 26$^{th}$ meeting of the General Congress of Weights and Measures (CGPM), it was decided to completely overhaul the SI system by defining seven constants from which all units of measurement were to be derived. The crucial point is that these constants are considered to have no MU. From then on, the Avogadro's *number* became Avogadro's *constant* which today is exactly:

$$N_A = 6.02214076 \times 10^{23}$$

Consequently, the molar mass of carbon 12 is no longer constant but has a standard uncertainty:

$$M(^{12}C) = 12.01074 \pm 0.00047 \text{ g/mol}$$

The molar mass of a molecule is determined in relation to the molar masses of each constitutive atom. For example, the empirical formula of theophylline is $C_7H_8N_4O_2$. The following table shows how to calculate its molar mass and the associated MU.

| Element | Atomic weight $M$ | $u(M_i)$ | $n_i$ | $n_i \times M_i$ | $u^2(M_i)$ | $n_i^2 \times u^2(M_i)$ |
|---|---|---|---|---|---|---|
| C | 12.0106 | 0.0006 | 7 | 84.0742 | $3.60\ 10^{-7}$ | $1.764\ 10^{-5}$ |
| N | 14.00686 | 0.00025 | 4 | 56.02744 | $6.25\ 10^{-8}$ | $1.000\ 10^{-6}$ |
| O | 15.9994 | 0.00021 | 2 | 31.9988 | $4.41\ 10^{-8}$ | $1.764\ 10^{-7}$ |
| H | 1.00798 | 0.00008 | 8 | 8.06384 | $6.40\ 10^{-9}$ | $4.096\ 10^{-7}$ |
| Total (g/mol) | | | | 180.1643 | | $19.226\ 10^{-6}$ |

The atomic weights and their standard uncertainties used for the calculation come from the Internet site of the Commission on Isotopic Abundances and Atomic Weights (CIAAW) managed by International Union for Pure and Applied Chemistry (IUPAC) [6]. They are normalized atomic weights established by considering an average isotopic composition. This explains the difference between the

carbon atomic weight used for calculation and the exact value given above because normalized carbon is a mixture of several isotopes.

The final molar mass of the molecule can thus be modified when it is marked. Since the total mass is an additive polynomial, the combined uncertainty is obtained by going through the standard variances of each constituent atom as explained in Section 6.7.1: this rationale is called the variance propagation law. If $n_i$ the number of $i$th atom in a given molecule, finally:

| | | |
|---|---|---|
| Total molar mass g/mol | $\sum n_i \times M_i$ | 180.1643 |
| Standard variance $u^2$ | $\sum n_i^2 \times u^2(M_i)$ | $19.226\ 10^{-6}$ |
| Expanded uncertainty $U$ g/mol | $2 \times \sqrt{u^2}$ | 0.0088 |
| Relative uncertainty *UR%* | | 0.0049% |

In the case of a calibrator solution obtained by weighting, the contribution of the molecule mass uncertainty to the overall MU is generally exceedingly modest compared to the other input quantities involved. For the molecular mass of theophylline, the relative uncertainty is only 0.0049%. But there are some special examples, such as gas analysis, where the contribution of molecular mass MU can be more important. However, the wider use of the mole as a unit in analytical science has two drawbacks:

- Its creation is recent, whereas the analytical sciences have a much older history and tradition to quantify a result.
- Unlike the meter or the kilogram when initially created, no concrete standards are available allowing an easy attachment to the mole. To have such standards, thousands or even millions of items would be needed, one for each chemical molecule. In a way, certified reference materials (CRM) play this role, but for a limited number of applications, as explained in Section 6.4.

Today most standards are dematerialized. For instance, the meter is no longer defined in reference to the prototype bar of platinum but as "the length of the path travelled by light in a vacuum in 1/299792458 of a second."

When the mole is defined in a comparable manner, many problems will be solved for analysts. However, the mole is not that often used in the laboratories, and it is frequent to express a concentration using other units, such as a mass (kg), a relative mass or a mass fraction (kg/kg or kg/l) which allows a connection to another more convenient SI unit, the kg. The downside of a relative concentration is that it is a dimensionless quantity. A widespread practice then – although discouraged by the BIPM – is to use a fraction to express a concentration, such as percent (or $10^{-2}$), ppm (part per million or $10^{-6}$) or ppb (part per billion or $10^{-9}$).

In solution chemistry, when it is necessary to consider the ionized form of a molecule, the Equivalent, noted Eq, is applied. It is a unit which integrates the electric charge and the mole, but which is not part of the SI. Thus, for the ions which carry a charge +1 or −1, like $Na^+$, $HCO_3^-$ or $Cl^-$, 1 mol = 1 Eq. For the ions of valence +2 or −2, as $Ca^{2+}$, 1 mol = 2 Eq. and so on for the other valence values.

From these units, various secondary units are derived, as well as fractions of units, such as the nanogram (ng), the millimole (mmol) or the milliequivalent (mEq). Finally, in some fields, such as biomedical analysis or microbiology, many analytes are poorly defined in terms of molecular structure, such as macromolecules, or mixtures of molecules. The solution adopted is to use the International Unit (IU) system, which is misnamed because it has nothing to do with the SI units. The definition of the IU differs from one substance to another.

It is the World Health Organization (WHO) Biological Standardization Committee that defines the IU of a substance based on the experimental measurement of its biological effects. For example, a 1000 IU tablet of oxytetracycline contains 1.149 mg of this antibiotic or 1.136 mg in its bi-hydrate form. A 1000 IU/ml preparation of vitamin D (a mixture of molecules with a vitamin D effect) contains 0.025 mg/ml.

## 6.4 Stage 1. Specify the Measurand

The four stages of the GUM general procedure are described in Sections 6.4–6.8.

Stage 1 starts with the specification of the measurand which is the "quantity intended to be measured" as explained in Chapter 1. In the analytical sciences, the term analyte is usually preferred to measurand. It is a chemical or biochemical species sought by the analyst and present in a matrix at known or unknown quantity. A measurement value of this quantity is generally the combination of:

- An operating procedure (i.e. a sequence of operations) that allows to prepare the sample by extracting the analyte from the rest of the matrix. It is done by playing on selectivity so as to obtain the analyte in such a form that it can be introduced into a measuring device without damaging it, or even to concentrate it to obtain a higher and detectable instrumental signal.
- And several input quantities, which are according to VIM "quantity that must be measured, or a quantity, the value of which can be otherwise obtained, to calculate a measured value of a measurand." They mainly represent an instrumental signal, a weight, a volume, or adjustment and correction coefficients that enter in the measurement model.

Situations are extremely diverse between non-destructive methods, indirect methods with calibration, primary methods, etc. The identification of the measurand consists precisely in describing and considering all these elements. As a first estimate of MU, the preanalytical operations of sampling will be neglected but fully addressed in a specific Section 8.3. As generally proposed in the guides designed for analysts, only analytical input quantities are taken into consideration.

The next chapter explains how to account for all the components of the MU in the case of a measurement obtained in a chemical or biological laboratory. With this simplified approach, the expression of a measurement is essentially based on the mathematical formula used to calculate the result. This is called the Measurement Model and consists of the mathematical relationship between the analyte and the various input quantities, such as reagent purities, the coefficients of the

calibration model, the volumes, and weights of the reagents, etc. Its general form is as follows:

General measurement model

$$Z = f(G_1, \ldots G_n, \ldots G_N) \tag{6.6}$$

$Z$ represents the output quantity and $f()$ any mathematical relationship; it is the analyte concentration which will be inferred from the information collected by the $N$ input quantities $G_1, \ldots G_n, \ldots G_N$ with $1 \leq n \leq N$. From a statistical point of view, the input quantities are the realizations of random variables with known probability distribution laws.

The example of a naive simple measurement model is given by the formula applied for calculating the concentration of the theophylline stock solution that was used to perform standard additions, as described in Section 5.2.2. To prepare such a solution, a known amount, $m$, of *analytical grade* theophylline is weighed on a laboratory scale. The purity $P$ of this reagent is provided by the manufacturer and is, in this case, 99%. This mass is diluted in a known volume $V$ of water. The measurement model which allows to obtain the concentration $X_0$ of the theophylline stock solution, is:

Measurement model for the stock solution

$$X_0 = \frac{m \times P}{V} \tag{6.7}$$

This measurement model contains three input quantities, i.e. $N = 3$. It allows us to establish clearly and without ambiguity what is going to be measured and represents the quantitative expression which links the value of the measurand to the three input quantities on which it depends.

At this stage, it is interesting to verify the consistency of the units. Thus, $X_0$ will evidently be expressed as mg/ml because $m$ is measured in mg, $V$ in ml, and quantity $P$ is dimensionless since it is a purity level. Later, for sample spiking, this stock solution is diluted and introduced such as the taken volume $V_p$ and the final dilution volume $V_f$ gives the expected final concentration as shown in this new model:

$$X_1 = \frac{X_0 \times V_p}{V_f}$$

But measurement models are often more complex and may combine many input quantities that are themselves other measurands. For example, the determination of lead by an ICP emission spectrophotometry with isotope dilution and coupling with a mass spectrometer or inductively coupled plasma-isotopic dilution-mass spectrometry (ICP-ID-MS) is an illustration of a complex measurement model [7]. In this case the model includes $N = 12$ input quantities:

Measurement model

$$Z = \left(\frac{Ms - Cs}{Pp.Wa}\right) \times \left(\frac{K.As6 - As8}{Ap8 - K.Ap6}\right) \tag{6.8}$$

Correction coefficient

$$K = \frac{\left(\frac{Ar8}{Ar6}\right)}{Rr} \times Rp \tag{6.9}$$

Table 6.1 contains the list of all input quantities with their definitions and units. It also provides an observed value $G_n$ for each that is used to compute an example lead concentration measurement value on a future CRM. It is important to keep all significant digits to make such a calculation. The observed measurement value is 1.999982505 mg/kg.

For this type of complex model, a preliminary precaution consists in checking what is called the dimensional equation. An error in the mathematical formulation can thus be corrected. In the case of Eq. (6.8), it is easy to check that the result of the dimensional equation is mg/kg, with most quantities being dimensionless ratios.

This approach to MU, based on the measurement model, is favoured by metrologists. However, the restriction to such measurement model is incomplete when classical analytical operations have to be considered, such as sampling in medical biology or sample preparation. Chapter 7 describes a more comprehensive method which we consider as better suited to analytical sciences and derived from the method accuracy profile (MAP).

**Table 6.1** LEAD – input quantities for the lead measurement model by ICP-ID-MS (CRM: certified reference material).

| Description of the input quantities | Symbol | Unit | $G_n$ |
|---|---|---|---|
| Mass of the spike | *Ms* | g | 0.7806 |
| Pb concentration in the spike | *Cs* | mg/kg | 0.4149 |
| Mass of the test portion | *Mp* | g | 0.4944 |
| Moisture (water activity) | *Wa* | % | 0.9255 |
| Isotopic abundance of $^{206}$Pb in the CRM | *Ar*6 | % | 40.0890 |
| Isotopic abundance of $^{208}$Pb in the CRM | *Ar*8 | % | 40.0954 |
| Ratio $^{208}$Pb/$^{206}$Pb in the CRM | *Rr* | % | 1.0189 |
| Ratio $^{208}$Pb/$^{206}$Pb in the test portion | *Rp* | % | 0.8994 |
| Isotopic abundance $^{206}$Pb in the spike | *As*6 | % | 0.9997 |
| Isotopic abundance $^{208}$Pb in the spike | *As*8 | % | 0.0001 |
| Isotopic abundance $^{206}$Pb in the test portion | *Ap*6 | % | 0.2454 |
| Isotopic abundance $^{208}$Pb in the test portion | *Ap*8 | % | 0.5290 |
| Output quantity | | | |
| Pb concentration in the test sample | *Z* | mg/kg | 1.999982505 |

There are also operating procedures where the analyte is defined by the method itself. This type of method is referred to as the "criteria method" by the Codex Alimentarius Commission [8]. The analyte can also be said to be defined *per se*, such as the determination of fiber or moisture in foods. Most often, the ultimate step consists of a series of weighings. They are gravimetric methods, and the uncertainty of the scales is used to calculate their combined MU.

It is frequent for traditional and long-standing standardized methods defined *per se* to be used as a reference to validate a new alternative method. This is the case when the so-called rapid microbiological counting method is used in food microbiology as an alternative to substitute for a traditional method using Petri dishes. The Association of Official Analytical Chemists (AOAC) has proposed to call this type of traditional reference method a *gold standard*. A common surprise is that the MU of the reference method may be higher than that of the alternative method. An example of a comparison of an alternative method to a reference method using the MAP and MU is presented in Section 10.2.2.

## 6.5 Stage 2. Identify Uncertainty Components

The second phase of the GUM procedure is crucial because it determines the rest of the process. From a practical point of view, it is advisable to use a cause-effect diagram or fishbone diagram to list the components (or sources) of the MU, identify their origin, and predict their influence on the overall uncertainty. It also avoids counting the same source multiple times and provides a simple mnemonic framework. This practical (and recommended) means is illustrated in Figure 6.2 for the ICP-ID-MS method of lead determination. All input quantities and the output quantities described in Table 6.1 appear in this diagram.

Each edge corresponds to an input quantity, considered as a potential source of uncertainty. It would be possible to go into more details. For example, all quantities involved in the preparation of Pb stock solution used for spikes at $C_s$ concentration could be added, such as volume, weights and lead purity corresponding to the model (6.7). Similarly, double weighing is used to quantify the moisture contents $W_a$.

However, in view of the considerable number of operations and quantities that lead to a chemical or biochemical measurement value, it is necessary to limit it to a

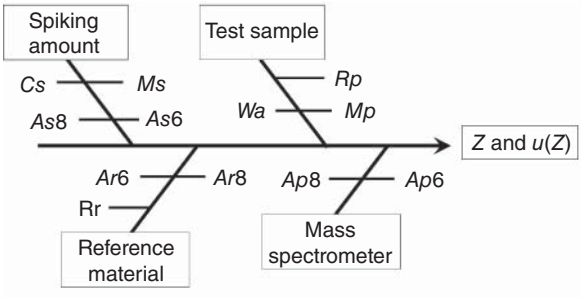



**Figure 6.2** LEAD – cause and effect diagram of the sources of uncertainty when determining lead by ICP-ID-MS.

reasonable level of detail. It is therefore a rather delicate stage, the main aim of which is to provide a perfectly clear and operational vision of what needs to be considered later.

The main problem with this approach is that the analyst has limited the measurement model to the reagents involved. Many other sources of uncertainty should be considered, such as extraction time, temperature, the aging of the reagents, and the skill level of the personnel. All these quantities are well known in the analytical science to have a definite influence on the result.

But the sources of variability, if they were to be included in the model in the same way as a volume or a weight, would pose serious modeling problems. At this stage, it seems difficult to put the various operations of mineralization, dry-ashing, or extraction into a mathematical equation! As claimed, in Chapter 7 a more empirical, comprehensive approach, sometimes called holistic, is proposed that is adapted to the practice of analysts and makes MU estimation simpler.

## 6.6 Stage 3. Quantify Uncertainty Sources

The third stage includes the following operations. For each previously identified component, it is required to quantify its amplitude of variation. In statistical terms, this means obtaining an estimate of the standard deviation that characterizes each random variable associated to an input quantity. However, if it seems impractical or cumbersome to quantify them experimentally one by one, the GUM recommends grouping the sources of uncertainty together, "as far as possible."

For example, if one identifies, on the one hand, variations in the setting of the measuring instrument, and, on the other hand, the aging of the reagents, as components to be evaluated, they can be grouped together by making measurements on different days, modifying the instrumental settings and the reagents from one day to another. GUM defines two ways to quantify component variability. This point is fundamental.

| | |
|---|---|
| Type A evaluation | It is empirical since it is based on statistical analysis of series of observations; some documents call it the *top-down* approach. |
| Type B evaluation | It is more deterministic, since it does not use observations but theoretical probability laws, chosen a priori; it is also called the *bottom-up* approach. |

From a practical point of view, and whatever the method of evaluation of each source (or grouping of sources), the aim is to obtain a statistical parameter of dispersion. Under pressure from statisticians, metrologists have given up calling the standard deviation this estimate of dispersion. Especially since, as explained in the next chapter, these parameters can be recombined, sometimes in a rather unconventional way. The term *standard uncertainty* instead of standard deviation has thus come into use.

By convention, the standard uncertainty of the input quantity $G$ is denoted $u(G)$. First, for each input quantity, it is essential to choose the most appropriate method

of evaluation of the standard uncertainty, either type A or type B. When several quantities are included in a measurement model, the combination of both types of evaluation is completely appropriate.

For example, in the case of lead determination by ICP-ID-MS, both modes of standard uncertainty evaluation were used. Thus, the $u(Ms)$ was estimated by a type B approach based on the probability law provided by the balance manufacturer, while $u(Wa)$ is the standard deviation of a series of repeated moisture measurements.

Next, at the fourth stage, the GUM describes how to combine the individual standard uncertainties of each source into an overall standard uncertainty called the combined standard uncertainty denoted $u_c(Z)$. In the rest of this text, the terms standard uncertainty and standard deviation may be indifferently used, with standard variance when it is a squared standard uncertainty.

### 6.6.1 Type A Approach

This approach starts by collecting observed measurement values purposefully obtained or from archive. The data that can be used to estimate a standard uncertainty are of various origins, and the following proposition list is not exhaustive.

#### 6.6.1.1 Accuracy Profile

We consider that MAP is probably the most suitable data source for an initial estimate of MU. Section 7.2 is entirely devoted to the details of the calculations from this type of data. In this case, the combined standard uncertainty can directly be obtained from the measurements made on the diverse validation materials.

An important constraint is that measurements must be collected according to a properly structured experimental design and not randomly. The rules to correctly organize the experimental design for collecting the suitable data to construct an accuracy profile are thoroughly explained in Section 5.4.2. The purpose of this organization is to trigger as many sources of uncertainty as possible and run the trials in separate series under intermediate precision conditions.

For example, if the replicates are scattered over several days, it is then possible to change the operators, the settings of the measuring system, the reagents, the calibration curve, etc. This strategy is only possible if the test material can be conserved long enough without undergoing a change considered too important, or prepared at will, for instance by weighing. If not, it will be necessary to resort to another organization of the replicates, which makes it possible to vary the sources of uncertainty that normally occur in the framework of routine usage of the method.

#### 6.6.1.2 Interlaboratory Study

If the laboratory takes part in an interlaboratory analysis, the collected measurements allow to estimate the MU of the test sample analyzed by all participants. It is recommended that the grand mean, estimated at the end of the study, be used as a reference value. The situation is even better if the analyzed material is a CRM. However, the standard deviation of reproducibility $s_R$ cannot be directly used

by the laboratory because is not representative of its own expertise, but over all laboratories.

The international standard ISO 21748 describes the safeguards to be taken in order to consider the laboratory's own performance. It explains how to deduce from the reproducibility standard deviation the laboratory-specific *intra-laboratory standard deviation* that approximates the intermediate precision standard deviation for each laboratory. The measurement model proposed in this standard is detailed in Section 7.2.1.

#### 6.6.1.3 Control Chart

A common idea is that, among all the data collected day after day in a laboratory, it should be possible to find a way to estimate the MU of the produced results. This may seem obvious, but several precautions are necessary before being sure it is possible. The main limiting factor is that repeated measurements must be performed on each quality control (QC) sample. While it is common in the industry it is unusual in laboratories.

Another challenge consists in grouping similar samples in terms of matrix, cautiously considering that the concentrations are close enough for the MU to be constant. In any case, this approach remains delicate and can easily lead to an overestimation of the uncertainty. On the other hand, the measurements obtained on the QCs that are used to establish control charts fulfill this constraint. This possible procedure is presented in Section 7.3.

#### 6.6.1.4 Proficiency Testing

For any accredited laboratory, proficiency testing is a compulsory interlaboratory comparison devoted to the verification of its ability to produce *accurate* results over time. In some cases, the proficiency testing organizer may use measurement values to estimate the test material MU (see Section 7.4). However, if derived MU incorporates all measurements and laboratories, it is impossible to obtain the dispersion interval characteristic of a given laboratory.

The material used in the proficiency testing can be converted into an internal reference material (IRM) applicable to a control chart. It is then possible to collect replicates under intermediate precision condition and have a specific MU estimate for the laboratory. This potential procedure goes back to the previous situation about control charts.

### 6.6.2 Type B Approach

This can be labeled as a theoretical approach since it can be applied without any experimental data. It is classically used for physical measurements but remains difficult for chemistry or biology. In its principle, it consists in associating to each uncertainty component a probability distribution law, chosen *a priori*. Standard deviations can easily be deduced from these theoretical distributions.

For example, the triangular distribution law is adapted for a measurand, such as digitized instrumental display. In modern instruments, the measured response is displayed as a set of digits on a screen, e.g. a spectrophotometer that measures the

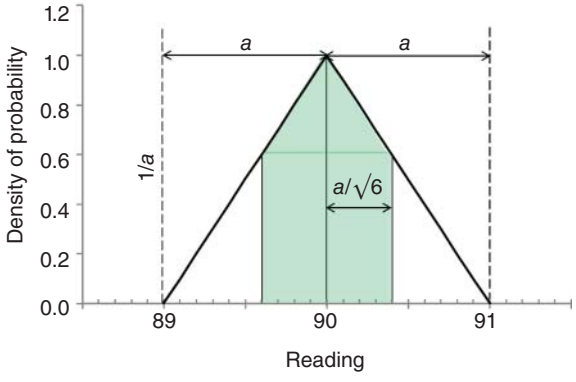



**Figure 6.3** Triangular distribution law applied to a digitized reading rounded to 90. The standard uncertainty is equal to $1/\sqrt{6} = 0.4082$.

optical density between 0 and 2500 milli absorbances would display a measurement value with four digits ranging from 0000 to 2500. The analog-to-digital converter included in the device could easily give more digits, but it was decided that the returned value is systematically rounded to the milli absorbance unit.

The model is simple: each reading corresponds to a continuous random variable $X$ bounded in the interval $\pm 1$ milli absorbance. The Normal distribution law varies between $\pm\infty$ and is consequently unsuitable. It is then considered that the probability of a reading linearly decreases when approaching the interval bounds. Classically, the triangular probability law is applied in order to model such an input quantity.

The Figure 6.3 shows a reading at 90 milli absorbances. True value is assumed to be located in the [89; 91] interval defined around the average which is thus worth 2 milli absorbances, traditionally denoted $2a$. The standard deviation of $X$ (or the standard uncertainty) is then simply deduced by posing:

$$u(X) = \frac{a}{\sqrt{6}} = \frac{1}{\sqrt{6}} = 0.4082$$

The *a priori* choice of the triangular distribution means that values far from the central value are less and less probable, while the central value is the most likely. Section 4.3 of GUM proposes several other types of theoretical distributions that can frequently be applied to estimate a standard uncertainty, such as the Normal distribution and the uniform distribution [1].

For the LEAD example introduced in Section 6.5 both types of approach are combined, depending on the information available, as shown in Table 6.2. Most of the time, information from manufacturers is used, such as the MU of the CRM or the scales. The key point is that it is possible to mix the two types of evaluation and that they are mutually combining.

## 6.7 Stage 4. Calculate Combined Uncertainty

### 6.7.1 Law of Propagation of Uncertainty

The combined uncertainty is a "standard MU that is obtained using the individual standard measurement uncertainties associated with the input quantities in a

**Table 6.2** LEAD – Approaches used in estimating standard uncertainty of input quantities for lead determination.

| Description of the input quantity | Symbol | Dimension | Approach |
|---|---|---|---|
| Mass of the spike | *Ms* | g | B |
| Pb concentration in the spike | *Cs* | mg/kg | A |
| Mass of the test portion | *Mp* | g | B |
| Moisture (water activity) | *Wa* | No | A |
| Isotopic abundance of $^{206}$Pb in the CRM | *Ar*6 | No | B |
| Isotopic abundance of $^{208}$Pb in the CRM | *Ar*8 | No | B |
| Ratio $^{208}$Pb/$^{206}$Pb in the CRM | *Rr* | No | A |
| Ratio $^{208}$Pb/$^{206}$Pb in the test portion | *Rp* | No | A |
| Isotopic abundance $^{206}$Pb in the spike | *As*6 | No | B |
| Isotopic abundance $^{208}$Pb in the spike | *As*8 | No | B |
| Isotopic abundance $^{206}$Pb in the test portion | *Ap*6 | No | A |
| Isotopic abundance $^{208}$Pb in the test portion | *Ap*8 | No | A |

CRM, certified reference material RM982.
Source: Adapted from Feinberg et al. [7].

measurement model." The starting point is the measurement model that can be identified in the general form:

General form of a measurement model

$$Z = f(G_1, \ldots G_n, \ldots G_N) \tag{6.10}$$

The previous two examples of measurement models show that function $f$ can take quite different forms and the number $N$ of input quantities $G_1, \ldots G_n, \ldots G_N$ can also be very variable. The goal of the fourth and final stage is to calculate the combined standard uncertainty of $Z$, i.e. to combine the various standard uncertainties estimated in stage 3 into a unique standard uncertainty. Three possible combination procedures are applicable depending on the mathematical form of the model.

#### 6.7.1.1 The Model Only Contains Additions and Subtractions

The classic property is applied: the variance of a random variable that is the sum of other random variables is the sum of their variances. This can be written:

Model

$$Z = G_1 + G_2 - G_3 \tag{6.11}$$

Standard variance

$$s_Z^2 = s_{G_1}^2 + s_{G_2}^2 + s_{G_3}^2 \tag{6.12}$$

Standard uncertainty

$$u_c(Z) = \sqrt{u^2(G_1) + u^2(G_2) + u^2(G_3)} \tag{6.13}$$

The well-known application of model (6.11) is when the result is expressed as the average of several replicates or as the sum of several intermediate measurements, e.g. vitamin A activity is computed by summing retinol and β-carotene concentrations. Since the standard uncertainty of each individual measurement is known, it is easy to derive the MU from the mean or the sum. This model is also used in Chapter 7 presenting practical estimation of the MU in different experimental situations.

#### 6.7.1.2 The Model Only Contains Products and Quotients

This situation is less frequent. In this case the squared coefficient of variation of the measurand is the sum of squared coefficients of variation of input quantities.

Model

$$Z = \frac{G_1 \times G_2}{G_3} \tag{6.14}$$

Standard deviation

$$\left(\frac{s_Z}{Z}\right)^2 = \left(\frac{s_{G_1}}{G_1}\right)^2 + \left(\frac{s_{G_1}}{G_2}\right)^2 + \left(\frac{s_{G_1}}{G_3}\right)^2 \tag{6.15}$$

Coefficient of variation

$$CV(Z) = \sqrt{CV(G_1)^2 + CV(G_2)^2 + CV(G_3)^2} \tag{6.16}$$

Relative standard uncertainty

$$\frac{u_c(Z)}{Z} = \sqrt{\left(\frac{u(G_1)}{G_1}\right)^2 + \left(\frac{u(G_2)}{G_2}\right)^2 + \left(\frac{u(G_3)}{G_3}\right)^2} \tag{6.17}$$

#### 6.7.1.3 The Model is a Complex Combination of Input Quantities

The measurement model for lead determination described by Eqs (6.8) and (6.9) is the example of more complex situation. It is no longer possible to have a simplified calculation formula. For this case, the GUM then proposes a very general approach based on the "law of propagation of uncertainty" derived from the Taylor series development.

At the end of stage 3, a set of standard uncertainties characterizing each input quantity in the measurement model is available. They are denoted $u(G_n)$. Starting from the general model of Eq. (6.10) which can mix any type of estimation approach, the combined standard variance of $Z$ is obtained combining the standard uncertainties of each input quantity according to the law of propagation of uncertainty. This leads to Eq. (6.18).

In the equation, $\frac{\partial Z}{\partial G_n}$ denotes the partial derivative of $Z$ with respect to $G_n$ and is named the sensitivity coefficient of the input quantity. On the other hand, $u(G_n, G_m)$ is the standard covariance between two input quantities.

$$u_c^2(Z) = \sum_{n=1}^{N} \left( \frac{\partial Z}{\partial G_n} \right)^2 u^2(G_n) + 2 \times \sum_{n,m=1\ n \neq m}^{N,M} \left( \frac{\partial Z}{\partial G_n} \right) \left( \frac{\partial Z}{\partial G_m} \right) u(G_n, G_m) \tag{6.18}$$

Sensitivity coefficient of

$$G_n c_n = \frac{\partial Z}{\partial G_n} \tag{6.19}$$

$$u_c^2(Z) = \sum_{n=1}^{N} c_n^2 \times u^2(G_n) + 2 \times \sum_{n,m=1\ n \neq m}^{N,M} c_n \times \mathrm{c}_m \times u(G_n, G_m) \tag{6.20}$$

The coefficient $c_n$ is called sensitivity coefficient because it indicates the relative part of the input quantity in the whole combined uncertainty. Using the sensitivity coefficients, it is possible to establish the uncertainty budget of the MU. It underlines the relative weight of each uncertainty component in the total uncertainty. The method to calculate the elements of the uncertainty budget is explained in a later section of this chapter. From a mathematical point of view, $c_n$ is a partial derivative of $Z$ with respect to $G_n$. Depending on the model, it may have no formal solution. In the next chapter it is explained how to manage complex models when partial derivatives are not computable or computed. A simplification of Eq. (6.18) consists in removing the covariance terms, i.e. assuming that the input quantities are uncorrelated.

$$2 \times \sum_{n,m=1\ n \neq m}^{N,M} c_n \times \mathrm{c}_m \times u(G_n, G_m) = 0$$

The adequacy of this assumption is based on the idea that input variables are perfectly independent. It is questionable but in practice the independence of random variables is sometimes difficult to ascertain. Moreover, special precautions are rarely established in the laboratory to ensure that measurements are perfectly independent, which would result in zero correlations.

Indeed, there are often many interconnecting elements in the analytical operating procedures, such as a single stock solution used to prepare various daughter solutions. This is a typical source of correlation between measurements which is experimentally difficult to quantify. Because this evaluation is difficult, even impossible, the proposed simplification is more based on a practical aspect than on scientific evidence. On the other hand, with respect to the various intermediate measurements that are used to calculate the end-result, this assumption seems quite correct. This finally gives the simplified model of Eq. (6.21).

$$u_c^2(Z) \approx \sum_{n=1}^{N} c_n^2 \times u^2(G_n) \tag{6.21}$$

### 6.7.2 Kragten Iterative Algorithm

Were Eq. (6.18) to be applied, it would be necessary to calculate the $N$ partial derivatives related to each input quantity. If the function which links the output and input variables is complex, this becomes very problematic, if not impossible. All functions do not have partial derivatives. GUM proposes different numerical methods to perform this operation without being obliged to go through the partial derivatives in an explicit form. For example, the Monte Carlo method and the Bootstrap algorithm are accepted alternatives [9].

The following proposed iterative algorithm is even simpler. The solution is based on discretization of the partial derivatives and can be developed in a worksheet. We will call it the Kragten's algorithm since it was first published by that statistician, member of the GUM committee [10]. In the original paper, the algorithm was described and applied to quite a simple model, i.e. the linear inverse prediction model, but it is adaptable to any measurement model.

However, if this is complex, it is better to use a somewhat more advanced programming language such as Python. Resource K is the application of the Kragten iterative algorithm to the complex LEAD measurement model described by Eqs. (6.8) and (6.9). Like other Python programs here, measurement values are recorded in the code but they could be read from an external file.

The starting point of the algorithm is to calculate the initial value of the output quantity, denoted $Z_0$using the measurement values of all input quantities. An iteration loop going through the $N$ input quantities is used to calculate $N$ modified values $Z_n^*$ by adding to each input quantity $G_n$ a slight specific variation $\Delta(G_n)$. To simplify things, the $\Delta(G_n)$ values are equal to one standard uncertainty $u(G_n)$ of each quantity. This means that all $u(G_n)$ are available as recommended at stage 3 of the GUM procedure.

At each step of the iteration loop, the new lead concentration is computed $Z_n^*$ with the modified value of one $G_n$. The difference between the initial value $Z_0$ and the new value $Z_n^*$ is calculated then squared. The sum of these $N$ squared differences approximately corresponds to the standard variance (Eq. 6.22).

$$u_c^2(Z) \approx \sum_{n=1}^{N} \left(Z_0 - Z_n^*\right)^2 \tag{6.22}$$

This algorithm is applied to the LEAD example and illustrated by Resource K.

**Resource K Iterative algorithm applied to LEAD (Python).**

```
import pandas as pd
import numpy as np
```

The `Pb_Concentration` function defines the measuring model of the LEAD application and must be modified for any other measurement model. Each input quantity is named and each corresponding observed values is transferred by the

*(Continued)*

means of an array called measure. This is the simplest way to do so that each input quantity can be iteratively modified.

```
def Pb_Concentration(measure):
  Ms = measure[0]
  Cs = measure[1]
  Mp = measure[2]
  Wa = measure[3]
  Ar6 = measure[4]
  Ar8 = measure[5]
  Rr = measure[6]
  Rp = measure[7]
  As6 = measure[8]
  As8 = measure[9]
  Ap6 = measure[10]
  Ap8 = measure[11]
  K = (Ar8 / Ar6) / Rr*Rp
  return(Ms*Cs / (Mp*Wa) * ((K*As6 – As8) / (Ap8 – (K*Ap6))))
```

A set of measured input quantities is used to illustrate the algorithm and compute the initial value. It must be changed if another measurement of lead is achieved.

```
Gn = np.array([0.7806,0.41495,0.4944,0.9255,40.089,
40.0954,1.0189,0.8994,0.99979,0.00013,0.2454,0.52903])
Gu = np.array([0.0002,0.0034,0.0002,0.00288675,0.0036,
0.00385,0.0041,0.0036,0.0000125,0.00001,0.001,0.0021])
```

The initial lead concentration is calculated by using the initial values and displayed.

```
Z0 = Pb_Concentration(Gn)
print(Z0)
dev_sqrt =[]
```

The loop of iterations takes one by one input values stored in `Gn` and add one standard uncertainty stored in `Gu`.

```
for n in range(len(Gn)):
```

The initial value of the input quantity $G_n$ is saved in a temporary variable to be restored after modification.

```
  in_old = Gn[n]
```

One standard uncertainty is added to the measure.

```
  In_new = Gn[n] + Gu[n]
  Gn[n]=in_new
```

The modified concentration is calculated.

```
  Z = Pb_Concentration(Gn)
```

The squared difference between initial and modified concentration is saved in an array.

```
Sqrt_dif = (Z -Z0)**2
dev_sqrt.append(sqrt_dif)
```

The initial value of the input quantity $G_i$ is restored.

```
Gn[n] = in_old
```

The loop is closed, and the uncertainty budget assessed.

```
Budget=[]
total_squares = sum(dev_sqrt)
for i in range(len(dev_sqrt)):
  budget.append(100*dev_sqrt[i]/total_squares)
```

Table 6.2 can thus be completed by the Table 6.3. The column $G_n$ contains the values used to calculate the initial value of $Z_0$. The column $u(G_n)$ lists the standard uncertainties. The initial value is $Z_0 = 1.999982505$ mg/kg. The modified results are obtained by iteratively adding to each input quantity, taken one by one, one standard uncertainty $u(G_n)$ which gives $G_n + u(G_n)$.

**Table 6.3** LEAD – Calculation of the combined standard uncertainty with the iterative algorithm.

| Step | | $G_n$ | $u(G_n)$ | Modified value $Z_n^*$ | Squared difference | Contribution (%) |
|---|---|---|---|---|---|---|
| 0 | | | | $Z_0 = 1.999982505$ | | |
| 1 | *Ms* | 0.7806 | 0.0002 | 2.00869 | $2.63 \times 10^{-7}$ | 0 |
| 2 | *Cs* | 0.41495 | 0.0034 | 2.02464 | $2.71 \times 10^{-4}$ | 30 |
| 3 | *Mp* | 0.4944 | 0.0002 | 2.00737 | $6.59 \times 10^{-7}$ | 0 |
| 4 | *Wa* | 0.9255 | 0.00288675 | 2.00194 | $3.90 \times 10^{-5}$ | 4 |
| 5 | *Ar*6 | 40.089 | 0.0036 | 2.00788 | $9.33 \times 10^{-8}$ | 0 |
| 6 | *Ar*8 | 40.0954 | 0.00385 | 2.00851 | $1.07 \times 10^{-7}$ | 0 |
| 7 | *Rr* | 1.0189 | 0.0041 | 1.99459 | $1.85 \times 10^{-4}$ | 21 |
| 8 | *Rp* | 0.8994 | 0.0036 | 2.02183 | $1.86 \times 10^{-4}$ | 21 |
| 9 | *As*6 | 0.99979 | 0.0000125 | 2.00821 | $6.31 \times 10^{-10}$ | 0 |
| 10 | *As*8 | 0.00013 | 0.00001 | 2.00816 | $5.18 \times 10^{-10}$ | 0 |
| 11 | *Ap*6 | 0.2454 | 0.001 | 2.01387 | $3.24 \times 10^{-5}$ | 4 |
| 12 | *Ap*8 | 0.52903 | 0.0021 | 1.99477 | $1.80 \times 10^{-4}$ | 20 |
| Combined standard variance: $u_c^2(Z)$ | | | | | $8.94 \times 10^{-4}$ | 100.00 |
| Combined standard uncertainty: $u_c(Z)$ | | | | | 0.02993 | |

For abbreviations see Table 6.2.
Source: Adapted from Kragten [10].

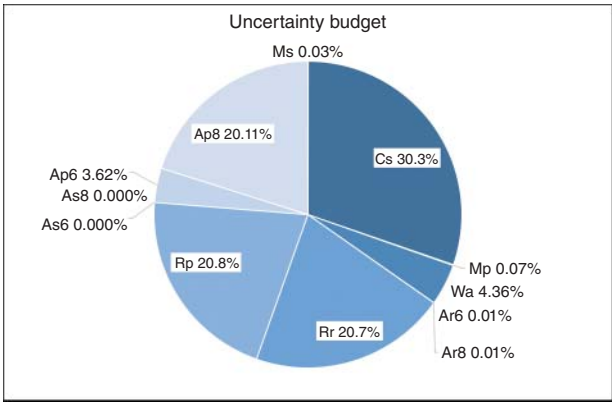



**Figure 6.4** LEAD – uncertainty budget.

The penultimate column contains the squared differences $\left(Z_0 - Z_n^*\right)^2$. As these differences can be ridiculously small numerically speaking, for these calculations it is necessary to use double precision variables. From these data, it is also possible to calculate the number of degrees of freedom associated with this estimate and use it to compute the coverage factor. This extension of the algorithm is not presented but available in the original publication. It was preferred to apply the standard GUM coverage factor for this example.

Most importantly, it is now possible to establish the uncertainty budget, described in the GUM as an indispensable complement to the MU estimate. The relative contribution of each quantity to the total combined uncertainty – its relative sensitivity coefficient – is obtained by calculating the ratio of the squared difference to the total standard deviation, as illustrated in Figure 6.4 and reported in the last column of Table 6.3. For example, the contribution of *Cs* (lead concentration in the spike) to combined uncertainty is:

$$100 \times \frac{2.71 \times 10^{-4}}{8.94 \times 10^{-4}} \approx 30\%$$

It can be concluded that the spiked concentration accounts for about 30% of the total MU and that three other critical input quantities are the isotopic abundance $^{208}$Pb in the test sample (*Ap*8), the ratios $^{208}$Pb/$^{206}$Pb in the CRM and in the test sample (*Rr* and *Rp*). When possible, better control of these quantities could help to reduce the lead MU. The graphical representation, e.g. in sectors, allows to better highlight the quantities that are influential on the MU.

The example illustrates the MU estimation for a single lead measurement value. If another sample is analyzed or the measure repeated, the whole calculation must be repeated with the new values of $G_n$. This solution is uneasy for contract laboratories where highly variable samples and matrices are seen daily. To avoid this limitation, the proposed solution described in Section 7.5 is to establish an uncertainty function applicable to any new sample and directly estimate MU of any inverse-predicted concentration.

There are dozens of commercially available uncertainty calculators that allow the recurrent calculation of MU. Some are even under free license, but many are commercial. The tricky point remains the input of the measurement model which may require some programming knowledge.

## 6.8 Calculate Expanded Uncertainty

The expanded uncertainty $U(Z)$ is simply "the product of a combined standard MU and a factor larger than the number one." This is then directly derived from standard uncertainty $u_c(Z)$ as indicated by formula (6.3) where $k_{GUM}$ represents the GUM standardized coverage factor, which depends on a coverage probability. Expanded uncertainty is the most practical form of MU.

Combined uncertainty is not easily understood by end-users because it remains theoretical like a standard deviation. To make it more understandable and practical it is multiplied by a coverage factor and expressed as an interval. For instance, it is well known that ±2 standard deviations include about 95% (precisely 97.7%) of the data if they are exactly distributed according to a Normal law. The coverage factor plays the same role and depends on the expected probability of the resulting coverage interval.

This is the most interesting expression of MU while a coverage interval is an "interval containing the set of true values of a measurand with a stated probability, based on the information available." Therefore, coverage interval can directly be used to express the analytical result:

Expanded uncertainty

$$U(Z) = k_{GUM} \times u_c(Z) \tag{6.3}$$

Coverage interval

$$Z \pm U(Z) \tag{6.5}$$

When the number of degrees of freedom ($df$) is accurately known at the end of the estimation of the MU, it can be used to have an exact value for the coverage factor. In most examples presented here, it is the quantile of the Student's $t$ law, for the given probability and the number of $df$. This point is developed more in Section 7.2.

When the information is not available, the standardized GUM coverage factor denoted $k_{GUM}$ is an interesting and easy solution. It can take simple conventional values:

| Value of $k_{GUM}$ | Coverage probability | Contents of the coverage interval |
|---|---|---|
| $k_{GUM} = 2$ | 95% | 95% possible true values of $Z$ |
| $k_{GUM} = 3$ | 99% | 99% possible true values of $Z$ |

For example, for the determination of lead by ICP-ID-MS, the test material contains $C_p = 1.99998$ mg/kg, and the combined standard uncertainty

$u_c(C_p) = 0.02993$ mg/kg as indicated in the bottom line of Table 6.3. The expanded 95% uncertainty is obtained by multiplying it by 2, which gives:

$$U(Z)_{95\%} = 2 \times 0.02993 = 0.05986 \text{ mg/kg}$$

And the relative uncertainty:

$$UR\%(\text{Z}) = \frac{0.05986}{1.99998} \times 100 = 2.99\%$$

The result can then be expressed as:

$$Z = 1.99998\ (\pm 0.05986) \text{ mg/kg}$$

This means that 95% of the true values, or considered as true values of $C_p$, are within the coverage interval [1.94012, 2.05984] mg/kg.

In most cases, the coverage probability is conventionally chosen to be 95%, i.e. $k_{GUM} = 2$. When the method of estimating the standard uncertainty makes it easy to obtain the real number of $df$, it is preferable to take the exact coverage factor, based on this number of $df$. In Section 7.2 the MU estimation based on the β-ETI is explained and the parameter $N_E$, called the number of effective measurements, is used to obtain the exact coverage factor and compared to the standardized.

## 6.9 Round the Result

The rounding of results is a very classic practice yet required by regulatory bodies. It is based on the concept of the number of significant figures. Above all, it is a simple and intuitive way of indicating the level of uncertainty that the experimenter gives to the measurements, and it is also a method to simplify their reading. Therefore, rounding can be considered as a method to express uncertainty. To perform result rounding it is thus possible to use its MU.

In some contexts, rounding is a regulatory requirement. For instance, the United States Pharmacopeia (USP) defined a standard operating procedure (SOP) for rounding analytical results. In this context, three different procedures are proposed and considered in agreement with three types of measurements: dissolution results, impurity and results close to limit of detection (LOD), and other results.

While using computers, it may seem superfluous to round results as far as the internal machine representation of a data contains a considerable number of figures. For example, an Excel worksheet uses a 64-bit floating-point internal format and generally stores data in 32 bits or more, though it is converted into an internal memory format, the number of significant digits can be extremely high. Above all, the data, such as it appears on the screen is not rounded in the sense that it is understood here but depends on the display format chosen by the user, whereas the calculations are done on all 32 bits or more.

The main disadvantage of rounding is that it necessarily generates a bias. It is therefore imperative to apply rounding only after all intermediate calculations have been completed. Many theoretical approaches have been published [8] but, according to ISO 17025 standard, the choice of the rounding method and the number of

significant digits is left to the laboratory, which can thus define its own rules [11]. Two simple rounding rules are applicable:

- When the first significant digit is between 5 and 9, the result is rounded to this decimal place and the MU will thus include one significant digit.
- When the first significant digit is <5, the result will be rounded to that next decimal place and the MU will therefore include two significant digits.

These rules can be translated into a mathematical formula, shown in Eq. (6.23). Let *Dn* be a decimal number with *ns* significant digits and **int**{ } the operator for obtaining the integer part of a decimal number. The following formula allows to determine *ns* by passing through the decimal logarithm of the number which is used to isolate the number of significant figures. The rounding is then done by considering *ns*.

$$ns = \mathbf{int}\{-\mathrm{Log}_{10}(Dn) + 1\} \tag{6.23}$$

**Resource L Rounding a result (Excel).**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Resource L: Rounding a result** | | | | | |
| 2 | | Raw data | Rounded data | Formula | | |
| 3 | Value | 1.999982505 | 2.00 | =ROUND(B3;INT(-LOG10(B4))+1) | | |
| 4 | Uncertainty | 0.059869686 | 0.06 | =ROUND(B4;INT(-LOG10(B4))+1) | | |
| 5 | | | | | | |
| 6 | **Examples** | | | | | |
| 7 | Z | U(Z) | Siginificant figures | Z* | U(Z)* | |
| 8 | 1.99998251 | 0.059869686 | 2 | 2.000 | 0.060 | |
| 9 | 75.23678 | 0.0278 | 2 | 75.240 | 0.030 | |
| 10 | 75.23678 | 0.00568 | 3 | 75.237 | 0.006 | |
| 11 | 75.23678 | 12.3689 | –1 | 80.000 | 10.000 | |

The Resource L worksheet illustrates this formula in Excel and is applicable to data such as the LEAD example. The raw result is in column `A`, the expanded uncertainty in column `B`, the rounded values in column `C`, and the Excel formula in column `D`. First, two internal functions are used `LOG10`, which provides the decimal logarithm, and `INT` which retains only the integer part. Finally the built-in function, `ROUND` completes the calculation using as argument the output of two functions. Finally, the Lead result can be expressed in various equivalent forms:

$$Z = 2.00 \pm 0.06 \text{ mg/kg}$$

$$Z = 2.00 \pm 3\% \text{ mg/kg}$$

$$Z = [1.94, 2.06] \text{ mg/kg}$$

## 6.10 Accuracy, Total Error, and Uncertainty

The introduction of the concepts of trueness, precision, and accuracy was mainly aiming to characterize a measuring method. Analysts are now familiar with these characteristics, although definitions may be fluctuating. On the other hand, MU is

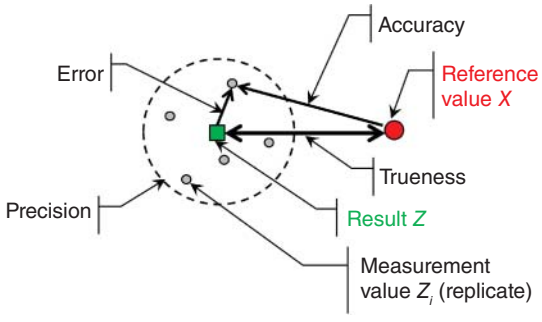



**Figure 6.5** Schematic representation of the main concepts used for method validation.

no longer focuses on the measurment method but on the analytical result, which can be described as the combination of the method, the laboratory, the type of matrix, the operator on a given day, and so on. To be explicit, the purpose of an interlaboratory analysis is to characterize the analytical method, using its standard deviation of reproducibility; it is not directly aimed to qualify the analytical result issued by a laboratory.

Figure 6.5 provides the schematic illustration of accuracy and the various characteristics or parameters used for method validation. According to the VIM definitions, trueness is the closeness of agreement between the average of an infinite number of replicate measurements and the reference value; it is a distance. The term "closeness of agreement" was preferred to the term "difference" to indicate this is a concept while several parameters are applicable to estimate trueness. In the figure, trueness is symbolized by a double-headed arrow.

Similarly, precision is defined as the closeness of agreement between replicate measures; it characterizes a dispersion. It is represented as a circle containing a known proportion of measurements (without going into the details of the calculation).

The accuracy, on the other hand, characterizes the closeness of agreement of one individual measurement to the reference value. Because these concepts are represented as vectors on the figure, it is easy to see that accuracy, as defined in the VIM, is the combination of trueness and precision.

MU is also estimated with diverse standard deviations or standard uncertainties, like the precision parameters. But the rationale is different and more complicated, as multiple sources of variation are accounted for, like in a jigsaw puzzle. The solution proposed by metrologists is to combine the various standard deviations corresponding to these sources.

It can be tedious, and so we propose something slightly different. It is also possible to use an intermediate precision standard deviation to cover most of the sources of uncertainty. This last remark highlights the fundamental difference between a reproducibility standard deviation and a combined standard uncertainty. Uncertainty can contain a standard deviation of precision but not the opposite.

A classic confusion consists in thinking that repeatability or reproducibility constitutes the uncertainty of measurement [12]. Some analysts believed that the

relative standard deviation of repeatability ($RSD_r$) is the appropriate estimate of relative uncertainty *UR*%. However, $RSD_r$ presents several major failings:

- It does not consider all the sources of variation included in the MU; therefore, it varies from day to day or from operator to operator.
- It does not consider the standard deviations of the measuring bias.
- It does not consider any coverage probability. This point is important as analysts often use the $RSD_r$ to characterize a method. In its usual expression, it corresponds to 1 standard deviation related to a given average value. At best, the coverage probability is 67% since the interval covers ±1 standard deviation. It would have to be associated with a coverage factor of 2 for the proportion covered to be 95%. For instance, when the $RSD_r$ is ±20% it means that 95% of the measurements are at ±40% around the announced result.

As already explained at the beginning of this chapter, in the field of clinical biology, the *TAE* has had some success. Like accuracy, it is based on a combination of trueness, using a bias, and precision introduced in the form of a standard deviation multiplied by a coverage factor to account the dispersion of measurements (formula 6.1) [13]. Figure 6.6 is a possible summarized comparison between *TAE* and MU. In the right part of the figure, the statements $\delta = 0$ and $u(\delta) \neq 0$ are not paradoxical but clearly underline two major principles of the GUM: the bias must be corrected before any MU estimation; and the bias standard uncertainty, if any, is included in the MU estimate.

In the conventional *TAE* model, for a given concentration level, the bias is considered as constant, this is the notion of systematic error, consequently its standard uncertainty is 0. Whereas in the MU model, the bias must be corrected and, because it is a random variable taking unpredictable values, it is possible to give an estimate of its standard uncertainty $u(\delta)$.

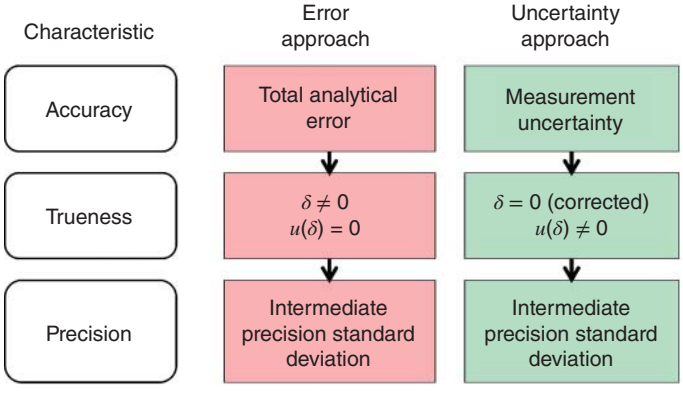



**Figure 6.6** Comparison between the total analytical error (*TAE*) model and the measurement uncertainty (MU) model.

## 6.11 Insights on Probability

Estimating the MU is an operation whose objective is to evaluate the *probable* dispersion of the true values of a measurand. Given that statistics is the science of probability, estimating the MU requires a statistical approach. Indeed, modern statistics appeared at the beginning of the nineteenth century, at the same time as the scientific way of thinking was being established.

Initially, famous mathematicians such as Pierre de Fermat (1607–1665), Blaise Pascal (1623–1662) or Thomas Bayes (1702–1761) elaborated the theory of probability calculation because of the interest in games and the haphazard nature of winning. And it is not a coincidence, if the Arabic word *az-zhar* indicates the game of dice, while the origin of the word random is more confusing. The concept of probability derives from the notion of randomness and is mentioned many times in this book and is rather delicate to grasp because it is not always intuitive.

A classic starting point is to refer to the game of heads or tails and the experiment conducted by the British mathematician John Edmund Kerrich while interned in Nazi-occupied Denmark in the 1940s. It consisted of tossing a coin $n = 10{,}000$ times and counting the number of times it landed on heads; this number being denoted $n_H$. In one experiment, the result was $n_H = 5067$. From this experiment, the concept of probability of having heads, denoted $\text{Prob}(H)$ can be defined, in a first meaning, as the relative frequency of the occurrence of heads: this leads to a possible definition of probability named frequential. That is, the ratio of the number of times heads was obtained to the total number of throws, denoted $n$, which gives:

$$\frac{n_H}{n} = \frac{5057}{10{,}000} = 0.5067 \text{ or } 50.67\%$$

This result is an observation of the real world. But intuitively it is easy to assume that the correct result should be equal to 0.5000 or 50.00% since a nonrigged or perfectly symmetrical coin has *theoretically* as much chance of falling on heads as on tails. By making this hypothesis, we unconsciously pass from the real world to an idealized world, posed *a priori*. To check if this passage is correct, it would require to be able to throw the coin an infinite number of times, which is of course unrealistic on the practical level, but which allows to propose a mathematical model, in the form:

$$\text{Prob}(H) = \lim_{n \to \infty} \frac{n_H}{n} = 0.5$$

It should also be noted that, when the experiment of 10,000-coin tosses is repeated, another value of $n_H$ would be obtained different from the first one. The real world is therefore not easily enclosed in a simple mathematical formula. When the number of possible values taken by a measure is known in advance, as in the toss of a coin where there are two possible values or the choice of a card in a deck of 52 cards, the notion of probability defined as a relative frequency is simple to grasp.

Things become more complicated with measurements that can take an infinite (or almost infinite) number of values, as a concentration measurement value obtained with a quantitative analytical method. The contribution of mathematics to solve this

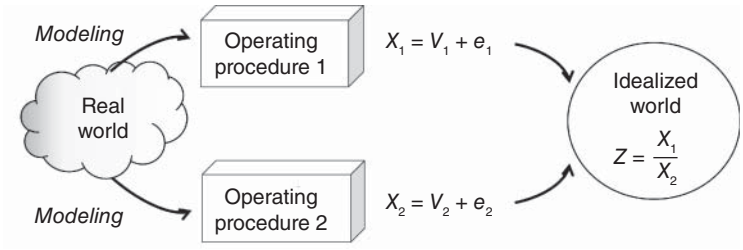



**Figure 6.7** Modeling: moving from the real world to idealized world.

problem has been to propose theoretical functions (or laws) that link a value to a probability density. They are known as Normal law, Student's law, binomial law, Fisher's law, etc. It is this additional level of abstraction that makes the notion of probability difficult to grasp because it is far from intuitive.

Figure 6.7 attempts to outline the approach extensively used by analysts to design an analytical procedure. As an example, consider the determination of fat in a food, expressed in g/kg of dry weight. This example is deliberately simple and serves only to illustrate the diagram. The first model consists in saying that lipids correspond to all substances extracted by an organic solvent, such as petroleum ether or hexane. This statement is often false, the model is more complicated and, depending on the food, pretreatments may be necessary before lipid extraction. We finally obtain a value of the lipid measurement denoted $X_1$ in the figure.

The second model is used to describe the measurement of the amount of dry matter present in the test sample, denoted $X_2$. Here again, various water elimination procedures are applicable.

However, behind these two models, which can be considered as chemical, are two statistical models which consist in declaring that each value $X_1$ or $X_2$ is made up of two quantities: a true concentration and a measurement error. These new models assume that in the sample to be analyzed, the lipids and dry matter are at constant and uniform concentrations, denoted respectively $V_1$ and $V_2$. There are also the errors made on each measurement, denoted here $e_1$ and $e_2$ which form the random part of the model and are more or less *probable*.

In the simple case illustrated here, it would be classic to say that these errors follow Normal probability distribution laws, with zero means. In the Section 8.3 which deals with sampling uncertainty, it is demonstrated that this idea of constant contents is also an ideal that never exists in the real world. Finally, the description of the real world will be reduced to ratioing $X_1$ and $X_2$ values.

As already pointed out, experimenters are facing a fundamental problem. When a measurement or an experiment is repeated, the obtained values or results are neither predictable nor identical. Statisticians therefore use the term "random variables" to describe this behavior of the measurement values.

The basic idea behind these few remarks is to remember that a value obtained at the end of an experiment, or an analytical procedure, is only a possible value, "more or less" probable. It is therefore important to keep in mind that the observed measures are only possible values and that repeating measurements would lead to other

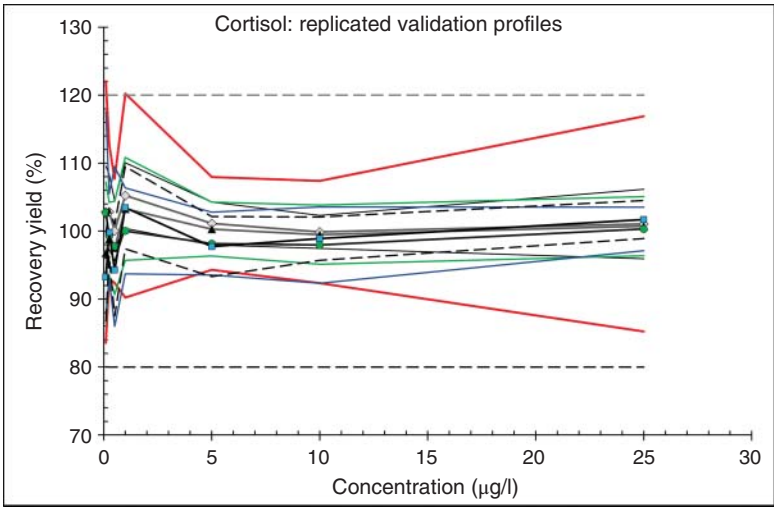



**Figure 6.8** CORTISOL – accuracy profiles from four validation studies.

values that are just as probable. The frequential definition of probability implies the idea of an infinite number of trials, so it is impossible to know exactly the probability associated with a value, that would then no longer be random.

On the other hand, the probability density functions developed by statisticians allow us to calculate, from a small number of observed values and for a given probability, a set of possible values contained, for example, in a prediction interval. To illustrate these remarks, Figure 6.8 puts together four repeated accuracy profiles obtained for the same method of analysis of CORTISOL by LC–MS–MS, over a substantial period, and without any modification of the analytical procedure [14].

The method for obtaining and interpreting accuracy profiles has been described in Chapter 5. In this example each profile consisted of three series of three replicates at seven concentration levels; thereafter, identical experimental design was repeated four times for cortisol. The dashed horizontal lines delineate the acceptance interval ±20% and all profiles as thin dashed lines.

The conclusion is that the method is valid over the validation range, whatever the chosen profile. On the same graphic, the β-γ-CTI of the obtained first profile is also reported as two thick solid red lines. It is interesting to remark that almost all β-ETIs are included within the β-γ-CTI obtained the first time. It brings some empirical proof that β-ETIs correctly predicts future measurements.

The construction of the relative uncertainty function from an accuracy profile will be described in Section 7.5.2. Figure 6.9 illustrates the relative uncertainty functions of the method. These functions are clearly affected by the random variability of the profile. For example, the relative uncertainty may vary from 5% to 12% at concentration of 1 mg/ml. Below this concentration, the differences are even greater, but at higher concentrations, the various functions converge. This observation of this

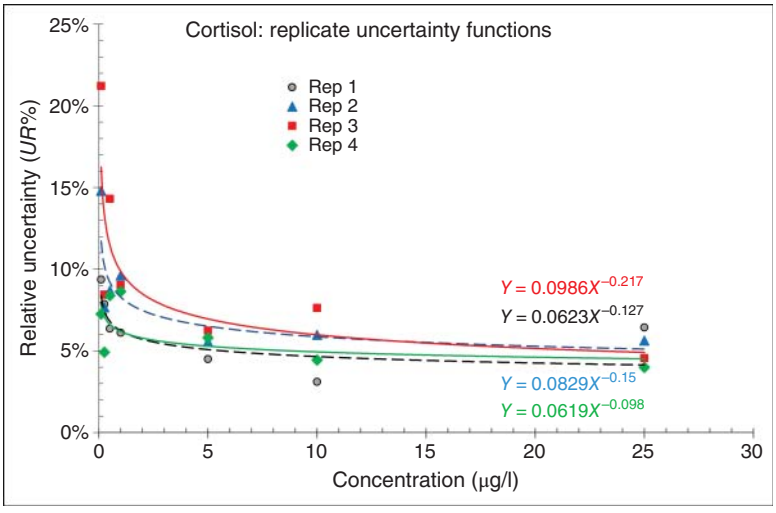



**Figure 6.9** CORTISOL – uncertainty functions of four replicated accuracy profiles, denoted Rep 1 to Rep 4.

variability of experimental data must be kept in mind as it is an illustration of the pending problems for the estimation of MU in analytical sciences.

Finally, since the beginning of the nineteenth century, the theoretical tools that are the distribution laws of probability, allow us to predict the probable behavior of the replicates of a measurement method. That is this approach, often called deterministic, that is presented here.

Since the 1970s, several new techniques, based on the intensive computing capabilities provided by computers, also allow the calculation of the probable dispersion of experimental measurements without using a theoretical law. They are sometimes called empirical or without-a-model. Unfortunately, they require different computational methods, such as the Bootstrap, which are not implemented in worksheets and require a specific program for each application [15]. Moreover, the regulatory authorities do not yet consider these innovative approaches as valid, for this reason they are not presented here, without presuming their importance and contribution to the field of analytical sciences in the coming years.

## References

1 Bureau International des Poids et Mesures (BIPM). (2012). International Vocabulary of Metrology — Basic and General Concepts and Associated Terms (VIM3), JCGM 200:2012, BIPM, Sèvres, France. https://www.bipm.org/ (accessed 23 July 2023).

2 Westgard, J.O. and Westgard, S.A. (2013). Total analytic error from concept to application. *Clinical Laboratory News* https://www.aacc.org/publications/cln/articles/2013/september/total-analytic-error (accessed 1 September 2023).

3 ICH International Council for Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human use (M10). (2023). *Harmonised guideline: bioanalytical method validation and study sample analysis*.

4 NCCLS 5 (2003). *Estimation of total analytical error for clinical laboratory methods; approved guideline*, NCCLS, document EP21-A.

5 Ellison, S.L.R. and Williams, A. (ed.) (2012). *Eurachem/CITAC Guide: Quantifying Uncertainty in Analytical Measurement*, 3e. ISBN 978-0-948926-30-3. www.eurachem.org (accessed 1 September 2023).

6 Commission on Isotopic Abundances and Atomic Weights. (2023). https://www.ciaaw.org (accessed 31 August 2023).

7 Feinberg, M., Montamat, M., Rivier, C. et al. (2002). Comparison of strategies to quantify uncertainty of lead measurements in biological tissue at mg $kg^{-1}$ level. *Accreditation and Quality Assurance* 7: 403–408.

8 Wilrich, P.T. (2005). Wilrich PT Rounding of measurement values or derived values. *Measurement* 37: 21–30.

9 BIPM, IEC, IFCC, ISO, IUPAC, IUPAP and OIML: *Evaluation of Measurement Data — Supplement 1 to the "Guide to the Expression of Uncertainty in Measurement" — Propagation of Distributions Using a Monte Carlo Method*, JCGM 101, (2008) https://www.bipm.org (accessed 1 September 2023).

10 Kragten, J. (1994). Calculating standard deviations and confidence intervals with a universally applicable worksheet technique. *Analyst* 119: 2161–2166.

11 Standard ISO 17025 (2017). *General Requirements for the Competence of Testing and Calibration Laboratories*. Genève: ISO.

12 De Bièvre, P. (2006). Accuracy versus uncertainty. *Accreditation and Quality Assurance* 10: 645–646.

13 Oosterhuis, W., Bayat, H., Armbruster, D. et al. (2018). The use of error and uncertainty methods in the medical laboratory. *Clinical Chemistry and Laboratory Medicine* 56 (2): 209–221.

14 Salamin, O., Ponzetto, F., Cauderay, M. et al. (2020). Development and validation of an UHPLC–MS/MS method for extended serum steroid profiling in female populations. *Bioanalysis* 12 (11): 753–768. https://doi.org/10.4155/bio-2020-0046.

15 Efron, B. and Tibshirani, R.J. (1993). *An introduction to the Bootstrap*. Chapman & Hall.

# 7

# Measurement Uncertainty in Analytical Sciences

## 7.1 Published Procedures: An Evaluation

The notion of measurement uncertainty (MU) has only recently emerged in analytical sciences. The 2005 version of ISO 17025 noted that "testing laboratories shall have and use procedures for estimating uncertainty of measurement. In certain cases, the nature of the test method may preclude rigorous, metrologically, and statistically valid, calculation of uncertainty of measurement" [1]. MU estimation is now mandatory for accredited laboratories and increasing in numerous application fields, such as environmental, toxicological, and clinical laboratories.

It quickly became apparent that the procedure classically used by other measurement professionals, in particular the Type B approach, was particularly difficult to apply to chemical or biochemical measurements. Since then, many proposals offering complementary, a priori better-adapted solutions have been published. Unfortunately, these proposals are very theoretical; likewise, validation procedures are often sectorial, linked to a specific field of application. It should be remembered, however, that Type A and Type B evaluations are not mutually exclusive and can even be complementary.

Figure 7.1 provides a graphical representation of several published procedures, and Table 7.1 associates the two-letter codes used at the bottom of the diagram with bibliographic references. These documents were selected because they present estimation procedures, official or consensus, specifically dedicated to the analytical sciences. The code AA corresponds to the guide to the expression of uncertainty in measurement (GUM) general procedure, already described in Section 6.2. The most striking example of the various available procedures is provided by the network of the European Directorate for the Quality of Medicines–Official Medicines Control Laboratory (EDQM-OMCL), where five different procedures are proposed for the so-called bottom-up approach only.

This list is not exhaustive, as new proposals are regularly published and/or updated. This results in two main observations:

– It is recognized that different procedures may give different estimates of MU. For instance, with the LEAD dataset, relative MU values ranging from 1 to 6 were obtained depending on the sources of uncertainty considered and the calculation method applied [7].

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

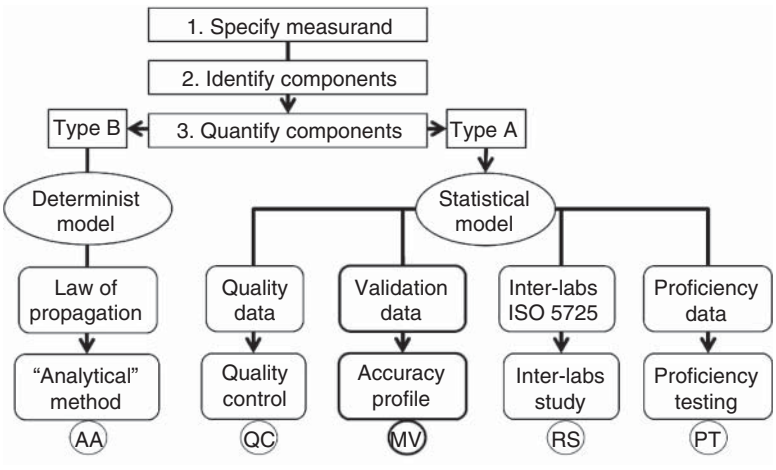



**Figure 7.1** Different MU estimation procedures proposed for the analytical sciences. The two-letter codes refer to the list of documents in Table 7.1.

**Table 7.1** Selected references of consensus methods for estimating measurement uncertainty in analytical sciences.

| Code | Organization | References | Years |
|---|---|---|---|
| AA | BIPM – GUM | [2] | 2012 |
| AA | EURACHEM-CITAC | [3] | 2012 |
| MV | SFSTP | [4] | 2017 |
| MV | AFNOR | [5] | 2010 |
| MV | EDQM-OMCL | [6, 7] | 2020 |
| MV | ISO | [8] | 2012 |
| PT | ISO | [9] | 2015 |
| PT | EDQM-OMCL | [10] | 2020 |
| QC | EDQM-OMCL | [11] | 2020 |
| QC | ISO | [8] | 2012 |
| RS | ISO | [12] | 2010 |
| RS | EDQM-OMCL | [13] | 2020 |

The organization acronyms are explained in Section 12.3.

– The direct consequence is that a harmonization effort promoted by analysts is crucial to propose the most judicious procedure or procedures, even if several approaches have already been the subject of recommendation or regulation.

Moreover, some specialized scientific publications describe dedicated applications. The journal *Accreditation and Quality Assurance* has published a few hundred examples for the analytical sciences. Even the vocabulary can be affected by these

discrepancies in some fields of application, such as medicines control, where two approaches are distinguished:

- *Top-down* to characterize the Type A evaluation based on data collected during interlaboratory or inter-comparison studies (designated by the code RS).
- *Bottom-up* is often based on a deterministic model and is synonymous with Type B evaluation.

The distinction between bottom-up and top-down does not shed much light on eventual harmonization [14]. Nevertheless some publications conclude that obtained values are nearly equivalent [15]. Several meaningful and suitable approaches are presented in the remainder of this chapter. They were selected because they allow reusing data that has already been collected for another purpose and do not require any new experimental effort. In that respect, the data already used for the method accuracy profile (MAP) is particularly appropriate. The main reason to recommend this procedure is that data collection follows strict rules fully compliant with MU estimation.

## 7.2 Use Method Accuracy Profile Data

### 7.2.1 Stage 1. Generic Measurement Model

To illustrate why we consider that MAP procedure can be an attractive approach to estimate MU, the different steps of general GUM procedure presented in Section 6.2 are inspected in detail, keeping in mind this specific application. Some publications present this extension of the accuracy profile to estimate MU, for example for a statin analytical method [16]. The first step is to formalize a measurement model that applies to all methods and avoids overly specific applications. The ISO/TS 21748:2017 standard can serve as a guide. Its scope covers two topics:

- MU evaluation is based on data obtained from interlaboratory studies organized in accordance with ISO 5725.
- The comparison of the MU, calculated from an interlaboratory study, with that obtained by applying the principles of the law of propagation of uncertainty.

In this standard, the proposed measurement model is generic and applicable to any field of measurement. According to the GUM recommendations, the calculation procedure can combine the Type A and B approaches. Before going any further, it should be remembered that MU characterizes a measurement made in a single laboratory and not the method applied in different laboratories, as is the case with reproducibility. It would therefore be nonsense to compare measurement uncertainty with a standard deviation of reproducibility [17]. As it is inconsistent to use the results of an interlaboratory analysis to estimate individual MU, several precautions are taken in ISO/TS 21748:2017 standard to avoid this problem.

When data are collected in a laboratory under intermediate precision conditions, there is no longer any contradiction in using the measurement model developed

in this standard, even though it was developed for interlaboratory comparisons. To obtain a MAP, it is mandatory to collect data under intermediate precision conditions, as explained in Section 5.4.2.

The similarity between the computation of reproducibility and intermediate precision variances was underlined in Section 3.2.1. In addition, since a full concentration range is covered with the accuracy profile, it will be possible to obtain MU values over the validation range at all concentration levels. Finally, an uncertainty function to calculate the MU for any concentration could be obtained. This function will be helpful in calculating the MU of any unknown sample result.

The ISO 21748:2017 measurement model principally considers a major part of the analytical process, i.e. the analytical operating procedure, and includes some pre-analytical steps. This applies to all analytical methods. It is presented in the form of the Eq. (7.1).

Generic measurement model

$$Z = X + \delta + B + \sum_{p=1}^{P} c_p G_p + E \tag{7.1}$$

where:

- $Z$ Recovered (or inverse-predicted) concentration of the analyzed material.
- $X$ Reference value assigned to the analyzed material (it is a constant).
- $\delta$ Intrinsic bias of the measurement method.
- $B$ Random variation is produced by the series factor effect that combines many sources of uncertainty.
- $G_p$ Other input quantities introduce a deviation from the nominal value of the reference $X$.
- $c_p$ Sensitivity coefficient of the input quantity (see Eq. 6.18).
- $E$ Residual random error, under repeatability condition.

The quantity $\sum_{p=1}^{P} c_p G_p$ represents the part of the measurement model that considers the slight deviations between the assigned value and the observed measurement values, e.g. during the preparation of the calibrators. As before, the sensitivity coefficient $c_p$ reflects the relative contribution of the quantity $G_p$ in the MU.

### 7.2.2 Stage 2. Generic Cause-to-Effect Diagram

In practice, the recommended starting point for listing MU components is to draw a cause-to-effect diagram, as described in Section 6.5. If approach B is applied, this is restricted to the input quantities involved in the formula applied for result expression. This limitation has been criticized, as it appears to ignore many sources of uncertainty that play a fundamental role in the measurements, such as sample preparation procedures or instrument settings. Table 7.2 presents a brief list of sources conventionally described in the literature in relation to sampling or sample preparation.

In the ISO 17025:2017 standard, there is an explicit list of critical points to be controlled in a laboratory to reduce the "variability," and the resulting uncertainty.

**Table 7.2** Sources of uncertainty in sampling and sample preparation.

| Sampling | Preparation of the sample |
|---|---|
| Heterogeneity (or inhomogeneity) | Homogenization and subsampling |
| Influence of the sampling strategy (random, stratified, proportional, etc.) | Drying |
| | Grinding |
| Sedimentation effects of bulk material | Dissolution |
| | Extraction |
| Bulk physical state (solid, liquid, gas) | Contamination |
| Effects of temperature and pressure | Derivatization (chemical effects) |
| Influence of sampling on the composition | Dilution errors |
| | Preconcentration and concentration |
| Transport and storage of the sample | Effects due to speciation |

Source: Adapted from EURACHEM-EUROLAB [18].

Through the reading of the different chapters, a list of practical requirements can be sorted out. It is the role of the inspector sent by the certification body, such as the French accreditation committee Comité français d'accréditation (COFRAC) or the American Association for Laboratory Accreditation (A2LA), to review this list and decide whether they are correctly implemented and controlled in the audited laboratory. Two main types of requirements can be identified corresponding to:

- The "commitment of results" where the laboratory is invited to participate in proficiency testing and demonstrate that it is capable of achieving acceptable results or scores.
- The "commitment of means" (or resources) where the laboratory must mobilize all human and technical aspects to ensure the performance, prudence, and diligence of its service at the declared level of quality.

These obligations give an insight into what must be considered as critical points for controlling the sources of MU. Several chapters of ISO 17025:2017 can be seen as participating in the catalog of MU components. It is adequate to organize them into a *generic* cause-to-effect diagram applicable to any analytical method issued from any laboratory, accredited or not.

Figure 7.2 represents a proposal summary to illustrate such a diagram. It differs from Figure 6.2 in that it identifies eight sources of uncertainty corresponding to eight chapters of ISO 17025:2017 [1]. It sometimes takes its name, Ishikawa diagram, from the Japanese quality promoter who proposed it in 1943. In an industrial context, the causes are categorized by the "5 M's" for machine, method, material, man/mind power, and measurement/medium. This diagram can also be called "5 M's".

As recommended, several sources of uncertainty can be clustered as one dotted area labeled *series effect* on the figure. This also appears in the measurement model of Eq. (7.1) as the $B$ random variable affecting the impact of series changes. It is intended to account for a majority of uncertainty sources that are present during the

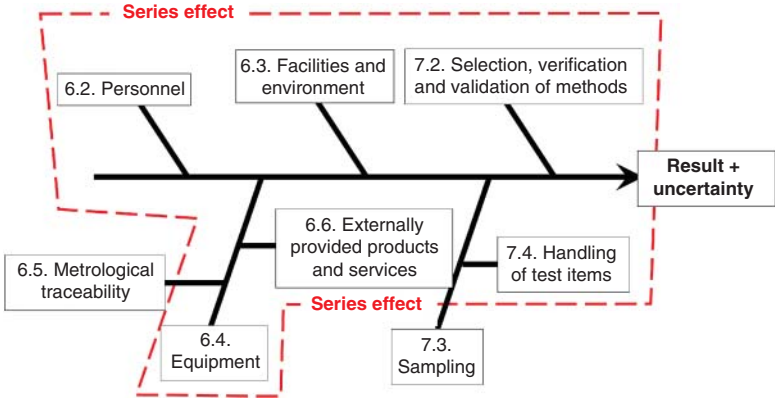



**Figure 7.2** Generic cause-to-effect diagram with eight main classic sources of uncertainty, according to [1]. The numbers in each box refer to the chapters of the standard. The dotted area groups sources into one series factor effect.

analytical part of the measurement process. To identify the standard uncertainty of this major part of MU, it is noted $u(Z_m)$ in the following, where subscript m stands for method. This observation ignores two other components of MU:

- $u(Z_s)$ the sampling standard uncertainty, which mainly characterizes the pre-analytical step and is discussed in Section 8.3.
- $u(Z_d)$ the so-called definitional uncertainty, which corresponds component of measurement uncertainty "resulting from the finite amount of detail in the definition of the measurand."

The comprehensive generic measurement model then takes the following form:

Comprehensive generic measurement model

$$Z = \left[ \underbrace{X + \delta + B + \sum_{p=1}^{P} c_p G_p + E}_{\text{Analytical process}} \right] + S + D \tag{7.2}$$

where $S$ and $D$ are random variables representing the random errors due to the sampling and the definition of the reference value $X$, respectively. After applying the law of propagation of uncertainty, the combined standard uncertainty of $Z$ thus becomes:

Combined uncertainty of $Z$

$$u_c(Z) = \sqrt{u^2(Z_m) + u^2(Z_s) + u^2(Z_d)} \tag{7.3}$$

In the first instance, the two standard uncertainties $u(Z_s)$ and $u(Z_d)$ are neglected to focus only on the standard uncertainty related to the method. This is realistic because, for many laboratories, the sampling (addressed in Section 8.3) is not part of their mission. As explained in Section 6.3 about the traceability to the SI units, the uncertainty on the reference value is negligible compared to the other sources.

### 7.2.3 Main Sources of Uncertainty in the Laboratory

To fully explain how this approach can be useful to quickly estimate the MU, it is necessary to review five major causes of Figure 7.2.

#### 7.2.3.1 Manpower

This is a delicate point since the aim is to assess the *competence* of the personnel responsible for the analyses. The ISO 17025 standard provides for the organization of training programs for the personnel, a qualification procedure for any new person entering the laboratory, and another for maintaining this competence. It is obvious that the implementation of a school-type examination, such as the analysis of material of known content, cannot be considered sufficient in a random context where acceptable deviations are not yet defined.

Sources of variability commonly attributed to the operators include systematic bias in the reading of a signal, weighting or volume error, and a different interpretation of a result, but all are correctable.

#### 7.2.3.2 Material and Handling of Items

These sources of uncertainty can generally be considered to be part of the pre-analytical phase. But, when samples are stored for a certain period of time prior to the analysis, the storage conditions may affect the results. The storage time and conditions should therefore be considered part of the MU. If sampling is the responsibility of the laboratory, the uncertainty associated with this operation must be evaluated. This is still currently under discussion, as it is difficult to achieve.

Some elements are presented in Section 8.3, and others are available in [18]. As soon as sampling is mentioned, many situations have to be considered, each of them being specific and requiring a particular statistical formulation: sampling plants in a field, blood in medical biology laboratories, or samples on a bulk lot of raw materials, etc. are all different in terms of statistical modeling. In addition, heterogeneity within a batch creates random variations between different samples that are not always easy to model.

Finally, the sampling procedure itself can produce a bias. For example, a water sampling device in a river creates a vortex that may displace some suspended particles.

#### 7.2.3.3 Method

This is probably where analysts best identify the sources of uncertainty. Most analytical procedures are a combination of operations, especially if the assay requires preparation. Without going into too much detail, it is necessary to sort out parameters that can globalize these sources of uncertainty. For example, the analyte recovery yield in a complex matrix can serve this purpose, even though the instrumental response may be affected by matrix effects. Analyte speciation can also increase these effects when a spike is used to estimate a recovery rate and the surrogate substance is in a chemical form different from the analyte.

Metrologists like to point out that often glassware material used at a different ambient temperature than that at which it was calibrated is a source of uncertainty. These effects are usually negligible, especially if the laboratory has air conditioners, essential for many modern measuring instruments. On the other hand, sample moisture can play a significant role if it is sensitive to possible hygrometric changes, which happens with biological samples that dry-freeze. There are also diverse sources of uncertainty due to calculations and signal processing. The selection of an approximate calibration model, early truncation and rounding of raw measurements, and incorrect settings of an integrator; all of these can lead to inaccuracies in the result.

The accuracy profile method can be used to detect these errors but not always to correct them. In trace analysis, blank or inherent concentration subtraction is significant. Uncertainty affects both the result and the timeliness of the blank correction. This problem is also encountered with the accuracy profile when standard additions are used, and it has not been possible to find a validation material without inherent content. Subtracting the so-called *no-addition* concentration also increases the MU (Chapter 1).

Finally, ISO 17025 cites an extreme situation where the analytical process is supposed to follow a certain stoichiometry, and the reaction proceeds in an incomplete way or with side reactions; then, it may be necessary to tolerate deviations from the expected stoichiometry.

#### 7.2.3.4 Machine/Equipment

Measurement equipment affects the MU, for example, the limited accuracy of the laboratory scales, the difference between indicated settings and actual temperature for a thermostat, the effects of contamination or memory for an automatic injection device, and many other technical downsides. That is why the instruments of an accredited laboratory must be regularly qualified and subject to regular maintenance contracts and control charts, all monitoring designed to reduce the risks of malfunction or drift.

For reagents, the concentration of a calibration solution is not known exactly. Many commercially available chemical compounds are not 100% pure and may contain isomers and mineral salts. The purity of these substances is usually provided by the manufacturer as a lower limit. Any assumption about the degree of purity introduces an element of uncertainty.

#### 7.2.3.5 Environment

The use of poorly designed spaces influences uncertainty. Some equipment requires air-conditioned rooms or clean rooms with positive pressure. The most striking example is that of microbiological methods, which require dedicated buildings known as *walk-in* rooms designed to avoid any cross-contamination in such a way that new samples never come in contact with already analyzed samples.

#### 7.2.3.6 Measurement and Other Sources

First, any statistical model must consider that there are random, unlisted, or overlooked effects, generally referred to as residuals, which contribute to uncertainty.

They are inevitably included in the list. There remain two sources of uncertainty that deserve special mention:

- Traceability of measurements. This issue is dealt with exhaustively by the Bureau international des poids et mesures (BIPM).
- Sampling uncertainty needs special attention and is discussed in Section 8.3.

### 7.2.4 Stages 3 and 4. Calculation of Combined Uncertainty

Looking at the diagram of Figure 7.2 it should be possible to propose an experimental method based on the multifactorial analysis of variance ($n$-way ANOVA) to estimate the standard uncertainty of each MU component. The latter will extract the variance component of each source in an equivalent way as between- and within-variances calculated in Section 3.2.

This statistical procedure is the generalization of the one-factor ANOVA. But it requires a substantial experimental effort, i.e. a large number of runs, as well as professional statistical software capable of handling complex models that may combine fixed-effect and random-effect factors. Therefore, only a limited number of publications are available to illustrate this procedure [19, 20].

However, at stage 3 of the GUM general procedure, it is judicious to aggregate the standard deviations of several identified components before evaluating them, especially when numerous components are present or difficult to estimate. In Figure 7.2, the dotted area exhibits together six of the eight sources of uncertainty. It is labeled as a *series effect* because it combines the method of analysis and the practical conditions of its application when the analysis is carried out.

Indeed, the same sample is reanalyzed over several days or in any other way that represents the so-called intermediate precision condition. All the sources of uncertainty grouped in the "series effect" will be brought into play simultaneously. In Section 5.4.2, the need to use an experimental design establishing intermediate precision conditions was pointed out for constructing the accuracy profile. Thus, the simplest and most practical approach consists in calculating an accuracy profile that can be used to estimate MU. It remains to be seen whether the estimation of the standard uncertainties and, ultimately, the combined standard uncertainty can be easily derived from these data.

This demonstration requires some simple mathematical operations. It starts with the generic measurement model of $Z_m$ described by Eq. (7.1), which is a simple additive polynomial relationship. After the law of propagation of uncertainty is applied, as shown in the Eqs. (6.11–6.13), the standard uncertainty of $u(Z_m)$ is given by the formula (7.4). For ease of calculation, the reference value $X$ is considered constant and known without uncertainty, in other words, $u(X) = 0$. Moreover, the quantities $\sum_{n=1}^{N} c_n^2 \times u^2(G_n)$ are assumed to be negligible compared to the other sources of uncertainty. Because both assumptions are plausible, this leads to the simplified formula of the Eq. (7.6).

Measurement model

$$Z_m = X + \delta + B + \sum_{n=1}^{N} c_n G_n + E \tag{7.1}$$

Standard uncertainty

$$u(Z_m) = \sqrt{s_B^2 + s_r^2 + u^2(\delta) + \sum_{n=1}^{N} c_n^2 \times u^2(G_n)} \tag{7.4}$$

There are two situations where these simplifications may lead to an underestimation of MU:

- Case 1. When a reference material certified or external reference material (CRM or ERM) is used. The standard uncertainty of the reference value $u(X)$ is provided by the manufacturer with the reference material when it is a CRM or by the proficiency testing organizer when it is an internal reference material (Section 7.4). This uncertainty is equivalent to the definitional uncertainty defined in the International Vocabulary of Metrology (VIM), $u(X) = u(Z_d)$.
- Case 2. When the standard addition method (SAM) is used. In this case, it is then mandatory to include in the measurement model (6.7) the consequence of the spiking process, which means keeping the quantity $\sum_{n=1}^{N} c_n^2 \times u^2(G_n)$ and estimating it.

But in most cases, after simplification, it becomes:
Intermediate precision variance (reminder)

$$s_{IP}^2 = s_B^2 + s_r^2 \tag{3.1}$$

Standard variance of $Z_m$

$$u^2(Z_m) \approx s_B^2 + s_r^2 + u^2(\delta)$$
$$u^2(Z_m) \approx s_{IP}^2 + u^2(\delta) \tag{7.5}$$

Standard uncertainty of $Z_m$

$$u(Z_m) \approx \sqrt{s_{IP}^2 + u^2(\delta)} \tag{7.6}$$

To demonstrate that, in the most general case, the data collected for an accuracy profile can also be used to estimate the analytical part of MU, it is satisfactory to show that formula (7.5) is equivalent to formula (5.7), i.e. that the variance $s_{TI}^2$ used to define the β-ETI around $Z$ is equivalent to the standard variance $u^2(Z_m)$. The quantities already presented in Section 5.3.1, relative to the β-ETI are used.

In formula (7.5), the variance of intermediate precision $s_{IP}^2$ is an estimate of the dispersion of the measurements under the influence of the "series effect" that integrates the major sources of the MU, as illustrated in Figure 7.2. This parameter can be estimated by an ANOVA and applying the formulas of (3.5) and (3.24). To obtain the standard variance of the bias $u^2(\delta)$ it is necessary to go back to pose the model of the bias:

Bias

$$\delta = \overline{Z} - X \tag{7.7}$$

Then, comes quite naturally the standard variance of the bias while $X$ is known without uncertainty:

Standard variance of the bias

$$u^2(\delta) = s^2_{\overline{Z}} + 0 = s^2_{\overline{Z}} \tag{7.8}$$

The standard variance of the bias is equal to the variance of the grand mean, whose formula corresponds to the Eq. (3.21).

Variance of the grand mean

$$s^2_{\overline{Z}} = s^2_{IP} \times \left(\frac{1}{IJQ}\right) \tag{3.21}$$

Standard variance of the bias

$$u^2(\delta) = s^2_{\overline{Z}} = s^2_{IP} \times \left(\frac{1}{IJQ}\right) \tag{7.9}$$

with:

$$Q = \frac{A+1}{J \times A + 1} \quad A = \frac{s^2_B}{s^2_r}$$

Finally:

Standard variance of $Z_m$

$$u^2(Z_m) \approx s^2_{IP} + s^2_{IP} \times \left(\frac{1}{IJQ}\right)$$

$$u^2(Z_m) \approx s^2_{IP} \times \left(1 + \frac{1}{IJQ}\right) \tag{7.10}$$

Standard uncertainty of $Z_m$

$$u(Z_m) \approx s_{IP} \times \sqrt{1 + \frac{1}{IJQ}}$$

Standard deviation of β-ETI

$$s_{TI} = s_{IP}\sqrt{1 + \frac{1}{IJQ}} \tag{5.7}$$

Standard uncertainty of $Z_m$

$$u(Z_m) \approx s_{TI}$$

The conclusion is clear, and the standard uncertainty, calculated for the generic measurement model, is approximately equal to the standard deviation used to calculate the β-ETI. This result can be explained by the fact that most of the sources of variation have been grouped into the "series effect." Furthermore, to construct the accuracy profile and calculate β-ETIs, it is mandatory to perform measurements according to the design of the experiment, which aims to alter the series effect and estimate the routine variability of the method. Then, to obtain a coverage interval, the coverage factor is set, allowing to determine the expanded uncertainty $U(Z)$. Two solutions are available for the coverage factor:

– Use the $k_{TI}$ which participates in the β-ETI and puts the value of $\beta\% = 95\%$, representing a 95% coverage probability. This is possible as the number of degrees of freedom is equal to $N_E$. This is the best solution and is recommended as the first choice, among others, by the GUM.

– If the actual number of degrees of freedom is unknown or difficult to compute, it is possible to take the simple $k_{GUM}$ standardized value proposed by the GUM, i.e. $k_{GUM} = 2$ for a coverage probability of 95% or $k_{GUM} = 3$ for a coverage probability of 99%.

When $N_E$ number of effective measurements is higher than 10, the value of $k_{TI}$ is remarkably close to 2, and the values of $U(Z)$ obtained by both solutions are then comparable. Therefore, in all cases where the number of degrees of freedom $N_E$ less than 10, the first solution is advocated. The direct Excel application of these equations is visible on lines 42–45 of the Resource H in Section 5.3.1. The calculation of the relative uncertainty *UR*% requires a few comments. By referring to Eq. (6.4), the equation should be:

$$UR\%(Z_m) = \frac{U(Z_m)}{Z_m} \times 100$$

where $Z_m$ is the grand mean of the inverse-predicted concentrations. Unless the analytical method is perfect, it is a biased estimate of $X$. Therefore, in this context, we recommend using the following formula where $X$ is the assigned value of the corresponding validation material:

$$UR\% = \frac{U(Z_m)}{X} \times 100$$

Because $Z_m$ is biased, it means that the bias would be included in the relative uncertainty and modify the estimation; a positive bias would diminish *UR*% and a negative bias would increase it. This point is an important difference between the total analytical error (TAE) concept and the MU, as underlined in Section 6.10 and illustrated in Figure 6.6. Therefore, in the following, the relative uncertainty is noted *UR*% to indicate it was calculated from the MAP.

A rapid example from THEOPHYlLLINE may illustrate the consequences of using $Z_m$ instead of $X$. For the first level where $X = 0.05$, the recovered grand mean $\overline{\overline{Z}} = Z_m = 0.0587$; consequently, the relative bias of about +17%. If *UR*% is calculated with $Z_m$, it gives 39.8% but 46.8% with $X$.

Finally, it can be asserted that the relative uncertainty better characterizes $X$. A short remark on this point is available in Section 3.2.2 and a more extended discussion in [21]. To illustrate the diverse calculations presented above, the THEOPHYLLINE dataset is used to give estimates for each validation material and compare the two proposed solutions for the coverage factor. Results are gathered in Table 7.3.

From this table, several conclusions can be drawn:

– For these examples, the GUM standardized coverage interval is always smaller than the coverage interval obtained by the exact method because the number of effective measurements is less than 10, except in one case; the standardized factor $k_{GUM}$ therefore tends to underestimate the expanded uncertainty.
– When the number of effective measurements tends towards the optimum (i.e. 11 in this example), the coverage factor $k_{IT}$ is close to $k_{GUM}$ The two intervals are almost identical because $N_E$ is optimal.

**Table 7.3** THEOPHYLLINE – estimates of expanded uncertainty using two methods of calculating the coverage factor.

| Concentration (µg/l) | $X$ | 0.05 | 0.10 | 0.50 | 1.00 | 2.50 | 10.00 |
|---|---|---|---|---|---|---|---|
| Grand mean | $\overline{\overline{Z}}$ | 0.0587 | 0.1115 | 0.5196 | 1.0013 | 2.5164 | 10.352 |
| Standard uncertainty | $u(Z)$ | 0.0117 | 0.0130 | 0.0350 | 0.0862 | 0.2749 | 0.5093 |
| Exact coverage factor | | | | | | | |
| Effective measures | $N_E$ | 7.01 | 9.59 | 7.02 | 5.69 | 10.91 | 9.22 |
| Probability β-ETI | $\beta$ | 95% | 95% | 95% | 95% | 95% | 95% |
| Coverage factor | $k_{IT}$ | 2.36 | 2.24 | 2.36 | 2.49 | 2.20 | 2.25 |
| Expanded uncertainty | $U(Z)$ | 0.028 | 0.029 | 0.083 | 0.214 | 0.606 | 1.148 |
| Relative uncertainty | $UR\%$ | 55.3% | 29.2% | 16.6% | 21.4% | 24.2% | 11.5% |
| Coverage interval | | 0.031 | 0.082 | 0.437 | 0.787 | 1.91 | 9.20 |
| | | 0.086 | 0.141 | 0.602 | 1.216 | 3.12 | 11.50 |
| GUM Standardized coverage factor | | | | | | | |
| Coverage factor | $k_{GUM}$ | 2 | 2 | 2 | 2 | 2 | 2 |
| Expanded uncertainty | $U(Z)$ | 0.023 | 0.026 | 0.070 | 0.172 | 0.550 | 1.019 |
| Relative uncertainty | $UR\%$ | 46.8% | 26.0% | 14.0% | 17.2% | 22.0% | 10.2% |
| Coverage interval | | 0.035 | 0.085 | 0.450 | 0.829 | 1.97 | 9.33 |
| | | 0.082 | 0.138 | 0.590 | 1.174 | 3.07 | 11.37 |

Figure 7.3 provides a graphical representation of the coverage intervals, which ultimately appear close. An interesting assumption is that the standard deviation of the β-ETI, $s_{IT}$, estimates most, if not all, of the combined standard uncertainty since it is limited to the analytical part of MU. The sources of uncertainty excluded by this approach are related to the sampling part and/or the validation material MU. The topic of sampling is discussed in Section 8.3.

Considering the reference values of the validation materials, other sources of uncertainty can be integrated into the model, especially in the case where certified reference materials are used. Sometimes the reference values are obtained through another method of analysis. This is quite common when it is possible to use a reference method, sometimes named *gold standard* method, as in the American literature.

## 7.3 Use Control Charts Data

### 7.3.1 Principles of the Shewhart Control Chart

The emergence of control charts in industry dates back to 1924 when W. A. Shewhart (1891–1967) introduced the first version of this quality control (QC) tool in the Bell Telephone Company, where he was employed. Since then, various

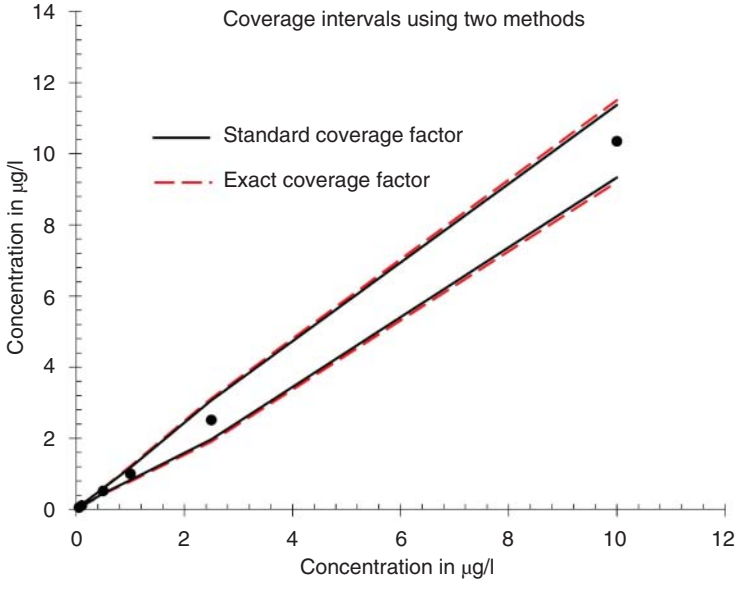



**Figure 7.3** THEOPHYLLINE – 95% coverage intervals. Dotted lines: tolerance interval method. Solid lines: standardized GUM.

control charts have been designed, each one adapted to a particular production context, as illustrated by the following (nonexhaustive) list, which consists of the ISO 7870 series of standards relating to control charts:

| Ref. | Year | Title |
|---|---|---|
| 7870-1 | 2020 | General guidelines |
| 7870-2 | 2013 | Shewhart control charts (replaces ISO 8258:1991) |
| 7870-3 | 2012 | Control cards for acceptance |
| 7870-4 | 2011 | Process adjustment control charts |
| 7870-5 | 2014 | Specific control cards |
| 7870-6 | 2016 | EWMA control charts (exponentially weighted moving average) |
| 7870-7 | 2020 | Multivariate control charts |
| 7870-8 | 2017 | Chart techniques for small series and small batches |
| 7870-9 | Project | Autocorrelated process control charts |

It is not our purpose to develop a complete theory about these various approaches, since it is available in many specialized books [22]. Our purpose is to show how to take advantage, *a posteriori*, of the data accumulated using a control chart to compile an estimate of MU. But several compilation methods are possible according to the type of chart.

This chapter is deliberately limited to the type most traditionally used in laboratories, namely the Shewhart control chart on the mean. Its principle is to

take a "sample of constant size at regular intervals." Unfortunately, the term *sample* remains highly ambiguous: for an analyst, a sample is a single object that must be analyzed; for a statistician, it is a set of entities forming a sample whose main characteristic is its size.

In other words, for the analyst, the principle of the Shewhart control chart implies, at a given interval of time, taking several test portions of one reference material and making replicates. In laboratories, such reference material is often referred to as QC sample. Let us note $J$ with $2 \leq j \leq J$ the number of replicates. This notation corresponds to the one used for the number of replicates per series of the accuracy profile experimental design.

In an analytical laboratory, the most conventional organization of a control chart consists in inserting at regular intervals QCs among routine samples forming a series. QCs are analyzed by the method, and the measurement values are graphically compared to a target value, denoted $T$ (for Target). Knowledge of the target value has, therefore, a major importance. As it is a question of trueness, Section 4.4 explains how the assigned *true* value $T$ of the QC reference material can be established and the warning and control limits used to verify that any new QC measure is not "out of specification" (OOS) and can be considered as acceptable. Different rules are proposed to make this decision.

When at least one new measurement value deviates from one of these limits, it is concluded that the analytical process presents unexpected conduct. In this case, various corrective actions must be taken before using the method for the following routine analysis. In the everyday vocabulary of the control chart, it is alleged that the process must remain under "statistical process control" (SPC) [22].

In the industrial context of standardized product manufacturing, such as a drug or a food, the target value $T$ can be an expected technical specification, such as the nominal contents of active ingredients, the packaging weight, and so on. For contract laboratories with a wide variety of samples, an internal reference material (IRM or ERM) can be used to set up a control chart because certified reference materials are unavailable or too expansive.

The reference value $T$, as well as the uncertainty, must have been predetermined in the same context of routine laboratory work. It is also popular to have simultaneously several control charts, and Figure 4.5 illustrates such a framework.

The dataset chosen for illustration is called ALBUMIN. It is available in an official document printed by the European Directorate for the Quality of Medicines and Healthcare (EDQM) devoted to the European medical and clinical laboratory network (OMCL Network of the Council of Europe [GEON]) [11]. It concerns the QC of purified commercial albumin solutions. Albumin solution mainly contains a unique monomer with a molecular weight of 66 kilodaltons (kDa), in addition, a variable percentage of polymers may be present, which are by-products occurring during fractionation, heating, and storage of the solution. Total polymer content (expressed in g/100 ml) is a quality criterion of this impurity, and a low measured value allows for good stability of the solution.

The high-performance liquid chromatography (HPLC) method for checking the total polymer content is based on exclusion chromatography (SEC) [23]. The

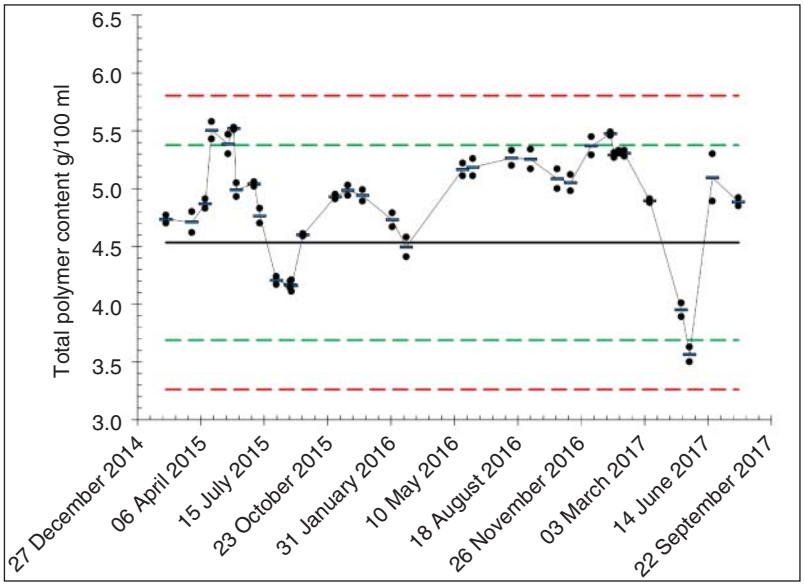



**Figure 7.4** ALBUMIN – control chart and QC. Warning limits are in green and control limits are in red.

impurity content of the reference material was determined exactly before starting. QC measures are denoted $Z_{ij}$ where $1 \leq i \leq I$ is the number of times, generally the number of days and/or each use of the method. From the number of replicates $J$ for each QC, three statistics are obtained: the mean, the range, and the standard deviation (equal to the range if $J = 2$). This leads to three types of charts. The most frequently encountered type, the Shewhart control chart, deals with the mean with $J$ small but greater than 1. If $J = 1$, another statistical processing must be applied, which will not be explained here.

QC measurements are plotted on a graph, i.e. Figure 7.4, where the interval between two QC is reported on the horizontal axis. On the vertical axis, the central line represents the assigned target value $T$ of the reference material. When using an IRM, this is usually the grand mean $\overline{\overline{Z}}$ of a set of replicates obtained during a preliminary study with the method to be controlled or a comparable alternative method. On the same axis, the following four limits, going from the top to the bottom, are added, having the general formulas, where $T$ is the target value:

Upper control limit

$$UCL = T + k_1 \frac{\hat{\sigma}}{\sqrt{J}} \tag{7.11}$$

Upper warning limit

$$UWL = T + k_2 \frac{\hat{\sigma}}{\sqrt{J}} \tag{7.12}$$

Lower warning limit

$$LWL = T - k_2 \frac{\hat{\sigma}}{\sqrt{J}} \tag{7.13}$$

Lower control limit

$$LCL = T - k_1 \frac{\hat{\sigma}}{\sqrt{J}} \tag{7.14}$$

The limits depend on a dispersion parameter, traditionally denoted $\hat{\sigma}$ and two coverage factors denoted $k_1$ and $k_2$. In Section 4.4, by convention, fixed factors were applied $k_1 = 3$ and $k_2 = 2$. Different methods are used to define $\hat{\sigma}$ and consequently, various control charts can be obtained:

| Method | $\hat{\sigma}$ values |
|---|---|
| Standard deviation of the preliminary study | $s$ |
| Standard deviation of repeatability of a preliminary study | $s_r$ |
| The known standard uncertainty of $T$ (applicable to CRM) | $u(T)$ |
| Predefined percentage $c$ of the target value | $T \times c$ |

The ALBUMIN example is based on the standard deviation of repeatability. The formula for calculating this standard deviation is presented in Section 3.2 and should be referred to for notations. In addition, Resource E is an Excel worksheet for calculating it. It is summarized as follows:

Standard deviation of repeatability

$$s_r = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(Z_{ij} - \overline{Z_i})^2}{I(J-1)}} = \sqrt{\frac{\sum_{i=1}^{I} SCE_i}{I(J-1)}} \tag{7.15}$$

Finally, the coverage factors can take on various values, but it is typically recommended to take:

| Type of limit | Coverage factor | Probability of coverage |
|---|---|---|
| Control | $k_1 = 3$ | 99% |
| Warning | $k_2 = 2$ | 95% |

### 7.3.2 Statistical Dispersion Intervals and Control Charts

Using the same dataset, it is possible to calculate the β-ETI, and β-γ-CTI described in Section 5.3 to quickly obtain a Shewhart control chart of the mean. The data may have been collected as part of a validation study or independently in a preliminary study with replicates. This allows us to get a MU estimate from QC measurement values.

Everything below is only applicable if the data collection method for the preliminary study is the same as it will be afterwards. This means that if the QC

**Table 7.4** ALBUMIN – simulated results of the preliminary study (total polymer contents expressed in g/100 ml).

| Series | Replicate 1 | Replicate 2 |
|---|---|---|
| Series 1 | 4.18 | 3.51 |
| Series 2 | 4.60 | 5.26 |
| Series 3 | 4.68 | 4.40 |
| Series 4 | 5.20 | 4.60 |
| Series 5 | 3.99 | 5.25 |
| Series 6 | 4.12 | 4.68 |
| Series 7 | 4.75 | 4.25 |
| Series 8 | 5.02 | 3.99 |
| Series 9 | 3.85 | 5.00 |
| Series 10 | 4.11 | 5.22 |

size is $J = 2$, the same number of replicates must be routinely collected. The control procedure in place amounts to analyzing size QC $J = 2$ at each date for the control of the method.

In the original document, the estimation of the target value is not presented. For pedagogical reasons, the results of a preliminary study were simulated using a normal law simulator available with Python. The study was supposed to last 10 days, $I = 10$ sets of $J = 2$ measurements whose values are collected in Table 7.4. The estimate of the target value was taken as $T = \overline{\overline{Z}}$ and the standard deviation of repeatability $\hat{\sigma} = s_r$.

As discussed in Section 5.4.4, the β-γ-content tolerance interval (β-γ-CTI) is considered as first choice to set the warning (*LWL* and *UWL*) and control (*LCL* and *UCL)* limits of the future control chart. The next step is to define the coverage probability $\beta$ and the confidence level $\gamma$. For this, the Food and Drug Administration (FDA) recommendations for biological assays could be considered [24]:

- The trueness should be in the ±20% range, which allows $\beta\% = 80\%$.
- Then, 2/3 (67%) of the QCs (i.e. 4/6) must fall between the warning limits, which gives $\gamma\% = 67\%$.
- Finally, as it is traditional, the control limits not to be exceeded must correspond to a confidence level $\gamma\% = 99\%$.

Finally, two β-γ-CTIs are computed having probabilities:

| | Content probability $\beta$ | Confidence level $\gamma$ |
|---|---|---|
| Warning limits | 0.80 | 0.67 |
| Control limits | 0.80 | 0.99 |

**Table 7.5** ALBUMIN – preliminary study.

| Parameter | | Warning | Control |
|---|---|---|---|
| Coverage probability | $\beta$ | 0.80 | 0.80 |
| Confidence level | $\gamma$ | 0.67 | 0.99 |
| Grand mean | $\overline{\overline{Z}} = T$ | 4.533 | |
| Std. dev. of repeatability | $s_r$ | 0.5954 | |
| Between-series standard deviation | $s_B$ | 0.0000 | |
| Standard deviation of β-γ-CTI | $s_{IC}$ | 0.6476 | |
| Coverage factor | $k_{IC}$ | 1.303 | |
| Lower limit | $L$ | 3.69 | 3.26 |
| Upper limit | $U$ | 5.38 | 5.80 |

Parameters of the β-γ-CTI to define limits of the control chart (g/100 ml).

The data from Table 7.4 were copied into Resource I, which provides the limits reported in Table 7.5. It is interesting to check whether the calculation of these limits, carried out classically, i.e. described in the standards and summarized by the Eqs. (7.11) and (7.14), give comparable results. For this purpose, the coverage factors $k_1 = 3$ and $k_2 = 2$, the mean as target value $T$ and, standard deviation of repeatability provided by Table 7.5 as an estimator $\hat{\sigma}$.

The standardized computation gives:

| Formula | Result |
|---|---|
| $LCL = T - k_1 \frac{\hat{\sigma}}{\sqrt{J}}$ | $LCL = 4.533 - 3 \times \frac{0.5954}{\sqrt{2}} = 3.27$ |
| $LWL = T - k_2 \frac{\hat{\sigma}}{\sqrt{J}}$ | $LWL = 4.533 - 2 \times \frac{0.5954}{\sqrt{2}} = 3.69$ |
| $UWL = T + k_2 \frac{\hat{\sigma}}{\sqrt{J}}$ | $UWL = 4.533 + 2 \times \frac{0.5954}{\sqrt{2}} = 5.37$ |
| $UCL = T + k_1 \frac{\hat{\sigma}}{\sqrt{J}}$ | $UCL = 4.533 + 3 \times \frac{0.5954}{\sqrt{2}} = 5.80$ |

With these new limits and the old ones rounded to two decimals, it is easy to check that similar values are obtained. They are plotted in Figure 7.4. This result is not surprising since the FDA recommendations are based on the same statistical principles as those that led to the development of the control charts. For the QC of commercial albumin solutions, the reference solution was used for three and a half years on 34 batches of bovine albumin. Figure 7.4 shows the Shewhart control chart obtained from these data with the limits previously calculated, while QC measurement values, with their dates, are recorded in Table 7.6.

The purpose here is not to comment on the QC problems of the analytical method exhibited by this chart but to explain how to use these 68 measurements to infer the MU of the reference solution and, subsequently, any sample close to that level.

**Table 7.6** ALBUMIN – QCs over two years expressed in g/100 ml.

| Date | Rep.1 | Rep.2 | Date | Rep.1 | Rep.2 | Date | Rep.1 | Rep.2 |
|---|---|---|---|---|---|---|---|---|
| 10 February 2015 | 4.70 | 4.77 | 27 August 2015 | 4.11 | 4.21 | 19 October 2016 | 5.17 | 5.00 |
| 23 March 2015 | 4.62 | 4.80 | 14 September 2015 | 4.61 | 4.59 | 9 November 2016 | 4.98 | 5.12 |
| 13 April 2015 | 4.83 | 4.91 | 4 November 2015 | 4.95 | 4.91 | 12 December 2016 | 5.29 | 5.45 |
| 23 April 2015 | 5.43 | 5.58 | 24 November 2015 | 4.94 | 5.03 | 11 January 2017 | 5.49 | 5.46 |
| 19 May 2015 | 5.30 | 5.47 | 17 December 2015 | 4.89 | 4.99 | 17 January 2017 | 5.31 | 5.27 |
| 28 May 2015 | 5.51 | 5.53 | 2 February 2016 | 4.67 | 4.79 | 24 January 2017 | 5.31 | 5.33 |
| 1 June 2015 | 4.93 | 5.05 | 24 February 2016 | 4.58 | 4.41 | 2 February 2017 | 5.33 | 5.28 |
| 29 June 2015 | 5.06 | 5.02 | 23 May 2016 | 5.11 | 5.22 | 14 March 2017 | 4.88 | 4.91 |
| 8 July 2015 | 4.70 | 4.83 | 8 June 2016 | 5.11 | 5.26 | 3 May 2017 | 4.01 | 3.89 |
| 3 August 2015 | 4.17 | 4.24 | 8 August 2016 | 5.20 | 5.33 | 16 May 2017 | 3.63 | 3.50 |
| 25 August 2015 | 4.14 | 4.20 | 7 September 2016 | 5.17 | 5.34 | 21 June 2017 | 5.30 | 4.89 |
| | | | | | | 1 August 2017 | 4.92 | 4.85 |

Source: Adapted from EDQM-OMCL [11], Table 2.1.

### 7.3.3 Estimation of the Reference Value Uncertainty

To obtain the MU of the mean value of the reference albumin solution, the proposal is to calculate the β-ETI with a coverage probability $\beta\% = 95\%$ so that it conforms with the GUM coverage factor. The Resource H Excel worksheet can do this computation after adapting it to the increased number of series. The question is to apply this method to 10 duplicates collected during the preliminary study or to the 34 series of routine QC.

Computed parameters for both datasets are summarised in Table 7.7. Some measures show that the method can be out of control and, therefore, significantly influence the MU estimate. Still, it is not a matter of eliminating them because they also contribute in the MU of the reference solution.

**Table 7.7** ALBUMIN – different estimation approaches for the MU of the reference albumin solution (g/100 ml).

| Parameter | Symbol | Preliminary study | Quality controls | Approach B |
|---|---|---|---|---|
| Number of data | $I \times J$ | 20 | 68 | 68 |
| Coverage probability | | 95% | 95% | 95% |
| Grand mean | $T$ | 4.53 | 4.91 | 4.91 |
| Std. dev. of repeatability | $s_r$ | 0.595 | 0.092 | |
| Between series std. dev. | $s_B$ | 0 | 0.462 | |
| Standard uncertainty | $u_c(T)$ | 0.610 | 0.478 | 0.467 |
| Number of *df* | $N_E$ | 18.9 | 34.3 | |
| Coverage factor | $k_{IT}$ | 2.09 | 2.03 | 2 |
| Coverage interval | | [3.26, 5.81] | [3.94, 5.88] | [3.97, 5.84] |
| Relative uncertainty | $UR\%(T)$ | 28.2% | 19.8% | 19.0% |

It is interesting to compare these results with those obtained by applying approach B and the law of propagation of uncertainty, as described in several guides reported in Table 7.1. To account for the fact that the number of replicates is 2 for each control and using the notations already employed, the combined standard uncertainty of the measurand is then:

$$u_c(T) = \sqrt{\frac{s_r^2}{2} + s_B^2}$$

$$u_c(T) = \sqrt{\frac{0.092^2}{2} + 0.462^2} = 0.467 \text{ g/100 ml}$$

The values obtained are presented in Table 7.7 and lead to close estimates, except for the preliminary study. The total number of measures has a strong influence on the MU estimation. It can be concluded that the relative uncertainty of the polymer content of an albumin solution with a concentration of about 4.9 g/100 ml is between ±19% and ±20%.

The use of β-ETI and β-γ-CTI presents the clear advantage of being simple and straightforward, on the one hand, to establish the parameters of a control chart, on the other hand, to estimate the measurement uncertainty of the reference material when the number of collected data is sufficient. After a certain number of QCs, it is also necessary to recalculate and modify the target value of the reference material and the limits of the control chart.

It is also possible to construct control charts at different concentration levels, allowing the relationship between MU and concentration to be estimated. One possible strategy for selecting the reference material is to exploit materials used during the validation procedure. They can be turned into reference materials for several control charts.

The measurements collected during the validation study can then be used to define the initial target values and control limits. Subsequently, regularly repeated QC measurements can refine the initial MU estimate. It is now well-established that MU tends to decrease as the method is used routinely and eventually stabilizes around a convergence value, provided that the sources of variability are constant over time.

## 7.4 Use Interlaboratory Comparison Data

The question of the different laboratory comparison procedures was already addressed in Sections 3.1 and 4.3, discussing precision and trueness, respectively. A proficiency testing scheme (PTS) is a procedure that consists of asking a large group of laboratories to perform a measurement, usually without replicates on a common sample, in most cases also without replicates. To verify their abilities in a specific field of application, laboratories can use their method. This procedure is a requirement for accreditation, and ISO 13528 international standard describes the statistical methods that can be used by proficiency testing organizers to interpret the results.

On the other hand, an interlaboratory study differs from a PTS because it is a study that brings together a group, often limited, of laboratories that are asked to perform replicates on the same sample using the same method. The aim is to evaluate the reproducibility of this method. This type of study could be mandatory to document the part of a standard related to the performance of a method. The ISO 5725 series of standards describes the organization and operation of this kind of interlaboratory analysis. However, the two comparison approaches may overlap, and previous descriptions may accept many extensions.

### 7.4.1 Proficiency Testing Scheme (PTS)

Because the participants do not perform replicates, it is impossible to estimate the respective MU for each result provided in a PTS. However, at the end of the test, the organizer can obtain an estimate of the value assigned to the test material, along with its standard uncertainty. For this purpose, chapter 7 of ISO 13528 defines an additive measurement model.

Unfortunately, there are also five ways to calculate the standard uncertainty of each model component. The drafters have therefore maintained a rather loose consensus, which illustrates the wide variety of areas in which proficiency testing is organized. These different methods of calculating the MU are specific to the activity of the PTS organizer. It may also be that some portions of the material used for the test are still available, or the organizer may suggest that samples be taken. It can then be used as an IRM for validation studies or the setting up control charts. In Section 8.3.2, the procedure for checking test material homogeneity, which described in Annex B of ISO 13528, is also used to assess the sampling uncertainty.

### 7.4.2 Interlaboratory Studies

When laboratories are required to make replicates while participating in an interlaboratory analysis, it seems logical to be able to calculate each result provided an estimate of its MU. ISO 21748:2017 describes a solution for this situation. This standard has already been mentioned in Section 7.2 to propose a generic measurement model because applicable to any analytical method.

Measurement model

$$Z = X + \delta + L + \sum_{p=1}^{P} c_p G_p + E \tag{7.1}$$

The definition of the elements of this model is slightly modified in the case of an interlaboratory analysis.

| | |
|---|---|
| $Z$ | Output quantity. |
| $X$ | Average value assigned to the test material analyzed. |
| $\delta$ | Intrinsic bias to the measurement method. |
| $L$ | Random variable for the laboratory factor effects and no longer the series effects. |
| $G_p$ | Deviations from the nominal value of the assigned value $X$. |
| $c_p$ | Sensitivity coefficient. |
| $E$ | Residual random error under repeatability conditions. |

The law of propagation of uncertainty explained in Section 6.7 applies to this additive model and, at the cost of some simplifications, the standard uncertainty of the measurand is equal:

Standard variance of $Z$

$$u^2(Z) \approx s_L^2 + s_r^2 + u^2(\delta) \tag{7.5}$$

Variance of the bias

$$u^2(\delta) = s^2_{\overline{\overline{Z}}} \tag{7.8}$$

Standard uncertainty of

$$\begin{aligned} Z\ u(Z) &\approx \sqrt{s_L^2 + s_r^2 + u^2(\delta)} \\ u(Z) &\approx \sqrt{s_R^2 + u^2(\delta)} \end{aligned} \tag{7.6}$$

An interesting proposal of the ISO 21748 standard is to introduce a new parameter called "intra-laboratory reproducibility." This ambiguous name is not only an oxymoron, but the idea behind it is interesting. It calculates an estimator of individual laboratory intermediate precision derived from the interlaboratory study data. The intra-laboratory reproducibility is denoted $s^2_{R_i}$, where $i$ is the laboratory number and $s^2_{r_i}$ the repeatability variance of the laboratory. Then it gives:

Laboratory repeatability variance

$$s^2_{r_i} = \frac{\sum_j^J (Z_{ij} - \overline{Z_i})^2}{J - 1} \tag{7.16}$$

**Table 7.8** LEAD – MU of the average of each laboratory.

| Labs | $\overline{Z_i}$ | $s^2_{r_i}$ | $s^2_{R_i}$ | $u(\overline{Z_i})$ | $UR\%(\overline{Z_i})$ (%) |
|---|---|---|---|---|---|
| Lab 01 | 2.03 | 0.001900 | 0.021349 | 0.14611 | 14.4 |
| Lab 02 | 1.94 | 0.003100 | 0.022549 | 0.15016 | 15.5 |
| Lab 03 | 2.17 | 0.060933 | 0.080382 | 0.28352 | 26.2 |
| Lab 04 | 2.43 | 0.006033 | 0.025482 | 0.15963 | 13.2 |
| Lab 05 | 1.92 | 0.000933 | 0.020382 | 0.14277 | 14.8 |
| Lab 06 | 1.88 | 0.000933 | 0.020382 | 0.14277 | 15.2 |
| Lab 07 | 2.02 | 0.001033 | 0.020482 | 0.14312 | 14.1 |
| Lab 08 | 2.02 | 0.003433 | 0.022882 | 0.15127 | 15.0 |
| Lab 09 | 2.09 | 0.003233 | 0.022682 | 0.15061 | 14.4 |
| Lab 10 | 1.99 | 0.000700 | 0.020149 | 0.14195 | 14.3 |
| Lab 11 | 2.02 | 0.000400 | 0.019849 | 0.14089 | 13.9 |

Laboratory average result

$$\overline{Z_i} = \frac{\sum_j^{n_i} Z_{ij}}{n_i}$$

Intra–laboratory reproducibility variance

$$s^2_{R_i} = s^2_{r_i} + s^2_L \tag{7.17}$$

Standard uncertainty of $\overline{Z_i}$

$$u(\overline{Z_i}) \approx \sqrt{s^2_{R_i}} \tag{7.18}$$

The formulas are applied to the LEAD dataset and recorded in Table 7.8. To verify the calculation, it must be remembered that $s^2_L = 0.01945$ for the balance design when no outlying laboratory is removed.

The graphic illustration of this interlaboratory MU estimate is provided in Figure 7.5a where the horizontal bars represent the laboratory mean, and the vertical error bars the coverage interval of each mean. The individual MU is somewhat important, while it was indicated in Section 3.4.2 that two laboratories were outliers. This remark underlines the leading role of the interlaboratory variance $s^2_L$ in the computation of the "intra-laboratory reproducibility variance" as it is occasionally 20 times higher than the individual repeatability variance.

When the outliers (data or laboratories) are removed, MU values are strongly reduced, as shown in Figure 7.5b. The scale on the *y*-axis is the same on both figures to show this dramatic modification. In this case, $s^2_L = 0.0003593$.

For a laboratory participating in an interlaboratory study, this approach is interesting as it is a good improvement of its work, but the MU obtained depends strongly on the performance of the other laboratories and must be interpreted with caution. It should be remembered that MU is specifically applicable to a measurement value obtained in one laboratory applying its operating procedure, whereas in this case, the estimation can be strongly influenced by other laboratories.

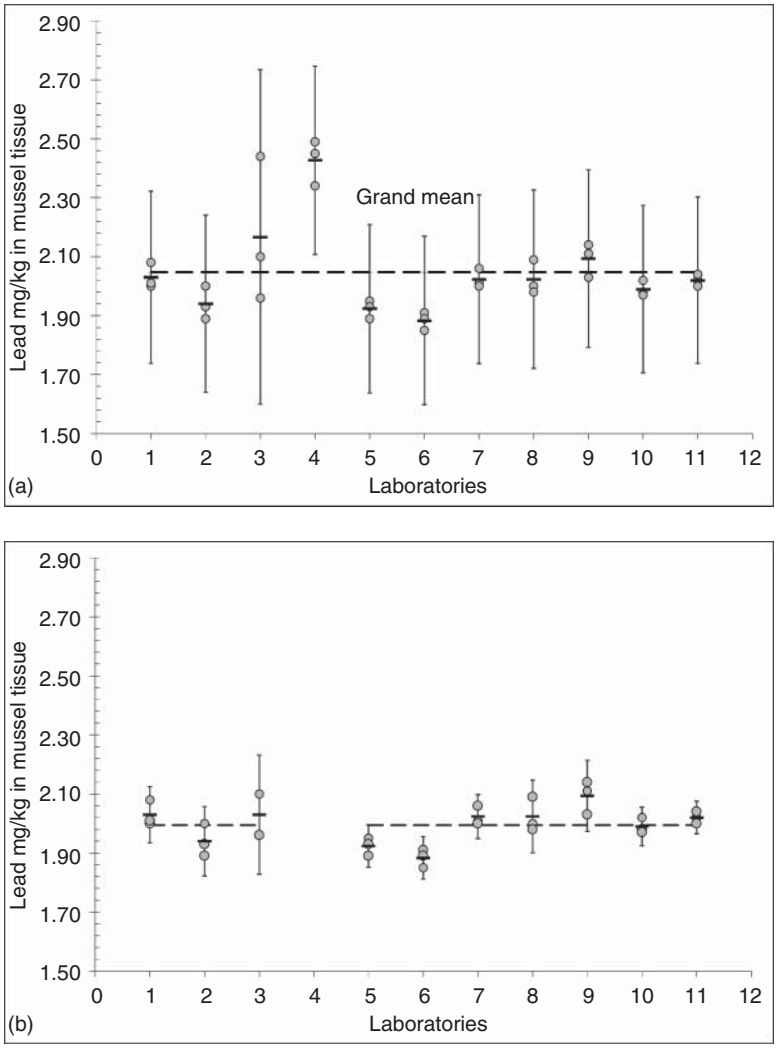



**Figure 7.5** (a) LEAD – laboratory coverage intervals including outliers, (b) LEAD – individual coverage intervals after removing outliers.

## 7.5 Uncertainty Functions

### 7.5.1 Horwitz's Model

It was previously explained that MU can vary depending on the instrument or measurement method. In the case of chemical or biochemical measurements, MU also varies, sometimes very strongly, depending on the concentration level. This variation is observable, regardless of the type of measurand or concentration, estimated by inverse prediction or assigned to a validation material. In some instances, this variation can be ignored if the sample is a product of a controlled industrial process because the concentration range of the analyte may be very narrow, and the associated MU can then be taken as invariant (see Section

Usually, once an MU estimate has been obtained for a given sample concentration, it is risky to extrapolate to another sample received by the laboratory unless the concentration is close. A possible method to avoid this disadvantage is determining the dominant relationship between MU and concentration, sometimes qualified as "functional relationship." Obviously, this requirement seems much less important for physical measurements where the same uncertainty applies to a wider range of measures.

Since the 1960s, interlaboratory studies and subsequent reproducibility calculations have been extensively used for quantitative method validation. By compiling thousands of these interlaboratory analyses, W. Horwitz demonstrated that reproducibility, i.e. precision, varies according to the concentration. The empirical model he proposed linked the relative standard deviation of reproducibility (also called the coefficient of variation of reproducibility), denoted $RSD_R$, to the concentration $X$. The basic rationale is that the $RSD_R$ could be multiplied by 2 every time the concentration is divided by 10 [25].

Equation (7.19) presents the initial mathematical formulation of this model published by Horwitz. The result is directly given as a percentage. $RSD_{TR}$ stands for theoretical relative standard deviation of reproducibility computed via the model while $RSD_R$ is the observed value. The empirical model relates concentration and precision in general, whatever the method operating procedure. It is, therefore, not directly applicable to the MU of any laboratory for any method; it is rather a very general model combining all types of analytical methods.

After rearrangement, an equivalent form is given by the Eq. (7.21), which links the theoretical standard deviation of reproducibility $s_{TR}$ to the concentration. It must be highlighted that this relationship is a power function of the same type as $f(Z) = aZ^b$. Moreover, it is a very convenient way to link these two quantities.

Horwitz model (theoretical relative standard deviation)

$$RSD_{TR} = \frac{s_{TR}}{X} \times 100 = 2^{1-0.5\,log_{10}(X)} \tag{7.19}$$

Alternative form of Horwitz model

$$RSD_{TR} = 2 \times X^{-0.155} \tag{7.20}$$

Theoretical standard deviation of reproducibility

$$s_{TR} = 0.02 \times X^{0.845} \tag{7.21}$$

Horwitz Ratio (*HorRat*)

$$HorRat = \frac{RSD_R}{RSD_{TR}} \tag{7.22}$$

A simple example illustrates the calculation of the $RSD_{TR}$. Consider a sample with a measured content of approximately 1 g/100 g. First, this concentration must be converted into International System units (SI), i.e. 1 g/100 g is equal to $10^{-2}$ kg/kg.

This value introduced into the equation gives the theoretical value of the $RSD_{TR}$. At this concentration level, $RSD_{TR}$ must usually be close to 4% as:

$$RSD_{TR} = 2^{1-0.5\log_{10}(10^{-2})} = 2^{1-0.5\times(-2)} = 2^{1+1} = 2^2 = 4\%$$

The remarkable success of this proposal led the official control bodies to create an acceptance criterion to check whether a standard deviation of reproducibility is acceptable or not. This is the *HorRat*, defined by the formula (7.22). This is the ratio of an observed relative standard deviation $RSD_R$ calculated at the end of an inter-laboratory study, and the theoretical $RSD_{TR}$ predicted by the Horwitz model. The proposed *HorRat* acceptance interval is [0.5, 2.0] [26].

Over time, this criterion and its acceptance have acquired a semi-official character which seems to give a kind of award to a method or, on the contrary, reject it. For example, the FDA, the Codex Alimentarius, and the European Union have adopted Horwitz's model to decide whether a method can be used for official control purposes, as explained later in this chapter. It can be expected that this decision rule is somewhat arbitrary. As remembered, the model is empirical and has no physical nor physicochemical basis that could justify a formal relationship between concentration and measurement precision. From an analytical point of view, one given method may be highly reproducible while another may not, but the main point is that both can fulfill the objectives assigned and are suitable for a given purpose.

Figure 7.6 illustrates Horwitz's model on a logarithmic scale ranging from 1 to $10^{-12}$ kg/kg or from 1 kg/kg to 1 ppt, as recommended by the author. A few values have special meaning for the customary laboratory discussion: $10^{-2,}$ which stands for 1%, $10^{-6}$ for part per million or ppm, $10^{-9}$ for part per billion or ppb, and $10^{-14}$ for part per trillion or ppt.

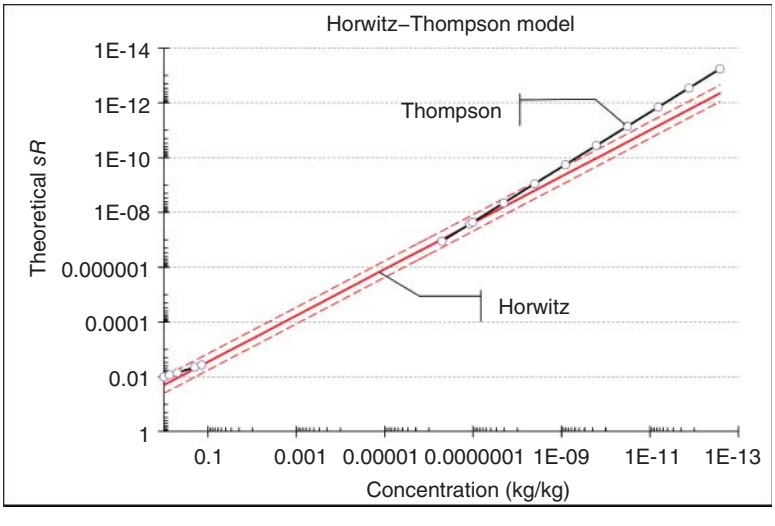



**Figure 7.6** Horwitz (solid line) and Thompson (dashed line) models. Both axes have logarithmic scales. Two thin dotted lines outline the *HorRat*.

A simple calculation shows that for 1.0 ppm, the theoretical standard deviation of reproducibility should be 0.17 ppm. If, abusively, it was assimilated to a standard uncertainty, then it is necessary to apply a coverage factor of 2.0, as explained in Section 6.10, and the expanded uncertainty would be 0.34 ppm or in a relative form 34%. However, this reasoning is somewhat fallacious. The reproducibility of a method is not the uncertainty of measurement [17].

It should also be noted that for concentrations below $10^{-10}$ kg/kg, or 0.01 ppm, the $RSD_{TR}$ is 50% or more. This means a dispersion of ±100% if the coverage factor of 2.0 is applied. At this concentration level, the values provided by the Horwitz model may become inapplicable for official control since they are between 0 and 2 ppm.

Let us consider a practical application. The Codex Alimentarius and FDA have set a maximum residue limit (MRL) for aflatoxin M1 in dairy products at 0.5 µg/kg, or $5 \times 10^{-10}$ kg/kg [27]. For this concentration, the Horwitz model gives the theoretical value $RSD_{TR} = 54\%$. In terms of dispersion of the measurements, this would be about 100%, and the sample acceptance range [0.0, 1.0] $10^{-10}$ kg/kg. However, this reasoning is somewhat fallacious but underlines some pratical application problems. The reproducibility of a method is not the uncertainty of a measurement.

Another model has been proposed as a set of three equations depending on the concentration [28]. The Horwitz-Thompson model is illustrated in Figure 7.6 by the large, black line.

| Concentration (kg/kg) | Standard deviation of reproducibility |
|---|---|
| $X < 1.2 \times 10^{-7}$ | $s_{TR} = 0.22 \times X$ |
| $1.2 \times 10^{-7} \leq X \leq 0.138$ | $s_{TR} = 0.02 \times X^{0.845}$ |
| $X > 0.138$ | $s_{TR} = 0.01 \times X^{0.5}$ |

The main disadvantage of the Horwitz–Thompson model is to be obtained from a wide variety of measurement methods. To have more specific method-oriented model, the revised version of ISO 5725-2 for interlaboratory studies includes a set of *functional relationships* between concentration and observed $s_R$ other than the Horwitz–Thompson model. In other words, empirical relationships between these two quantities. Four models, numbered from I to IV, are proposed to relate the standard deviation of reproducibility $s_R$ to $X$, as summarized in Table 7.9.

These relationships allow more realistic models to be fitted and adapted to each analytical technique. Similar models describe the standard deviation of repeatability. Unfortunately, ISO 5725-2:2019 does not specify how to choose the *best* relationship; there is no acceptance or quality criterion, as explained by *HorRat*. A major part of regulatory guidelines uses percentages to define parameter acceptance limits. For this reason, the Horwitz model was quickly adopted and became a suitable reference for the FDA, the Codex Alimentarius or the European Union, and many others [24, 29, 30].

Besides the acceptance limits for precision, acceptance intervals for trueness were added, as illustrated in Table 7.10. This is a composite outline of several documents,

**Table 7.9** Models relating $s_R$ to $Z$ proposed in ISO 5725-2:2019.

| Model | Type | Model | |
|---|---|---|---|
| I | Linear through 0 | $s_R = bZ$ | (7.23) |
| II | Linear | $s_R = a + bZ$ | (7.24) |
| III | Quadratic | $s_R^2 = a^2 + (bZ)^2$ | (7.25) |
| IV | Power | $s_R = aZ^b$ | (7.26) |
| | Linearized form | $\log(s_R) = \log(a) + b.\log(Z)$ | |

**Table 7.10** Trueness and precision acceptance limits applied in several official regulatory guidelines.

| | | Recovery yield (%) | | Reproducibility RSD (%) | |
|---|---|---|---|---|---|
| Concentration (kg/kg) | Unit | Lower (%) | Upper (%) | Expected (%) | Horwitz (%) |
| 1 | 100% | 98 | 102 | 1.3 | 2 |
| $10^{-1}$ | 10% | 98 | 102 | 1.9 | 3 |
| $10^{-2}$ | 1% | 97 | 103 | 2.7 | 4 |
| $10^{-3}$ | 1‰ | 95 | 105 | 3.7 | 6 |
| $10^{-4}$ | 100 ppm | 90 | 107 | 5.3 | 8 |
| $10^{-5}$ | 10 ppm | 80 | 110 | 7.3 | 11 |
| $10^{-6}$ | 1 ppm | 80 | 110 | 11.0 | 16 |
| $10^{-7}$ | 100 ppb | 80 | 110 | 15.0 | 22 |
| $10^{-8}$ | 10 ppb | 60 | 115 | 21.0 | 32 |
| $10^{-9}$ | 1 ppb | 40 | 120 | 30.0 | 45 |

and more precise values can be found in official documents. The variation range of concentration is exceptionally large and expressed power of 10 for kg/kg, as in Figure 7.6. The second column gives the informal name of the concentration units often used in practice. Figure 7.7 is a graphical illustration of this table. The horizontal axis is generally expressed in logarithms to cover a broad range of concentrations. This indicates the leading role of concentration in the definition of relevant acceptance limits.

As explained in Section 4.1.1, the recovery yield is considered by regulators as the most convenient parameter to express trueness. It is noteworthy that all acceptance criteria are expressed as percentages. This practice is quite common in analytical sciences, but is not without drawbacks. Just like any ratio, it is a combination of two values. In the present context, both have the same units, resulting in a dimensionless parameter. Consequently, the initial scaling is lost, e.g. one does not know if the value is obtained at ppb or ppm level.

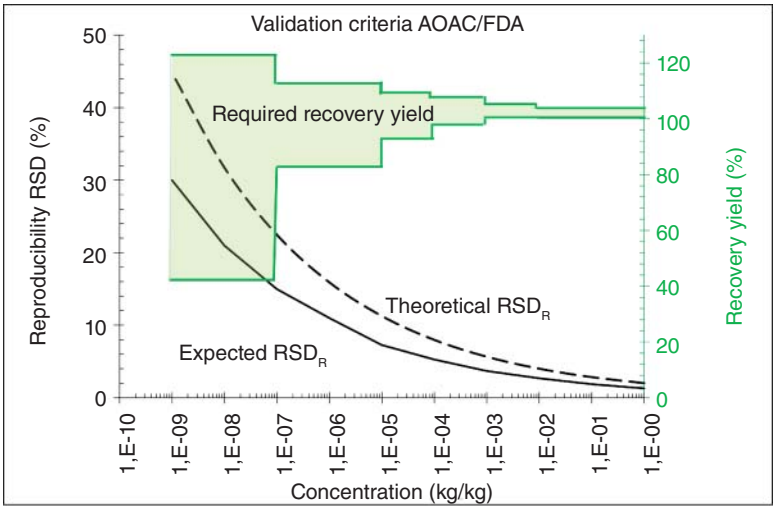



**Figure 7.7** Precision and trueness acceptance criteria proposed in various official guidelines.

### 7.5.2 Fitting the Uncertainty Function

It was previously emphasized that the Horwitz and other models relating reproducibility and concentration cannot be directly used to predict MU at any concentration but only suggest that a functional relationship does exist between measurement dispersion and concentration. This relationship is hereafter referred to as the uncertainty function of measurement. Having such a function is interesting for the routine use of the method because it will allow us to associate an MU with any unknown sample. The only limitation is the concentration of the sample within the range where this function has been determined.

For example, if one relies on the data of an accuracy profile, the uncertainty function will only apply in the validated range. Table 7.11 brings together the main uncertainty functions encountered in the literature. They are presented in two forms deductible from each other, either the standard uncertainty $u(Z)$ or the relative uncertainty $UR\%$, respectively; this link is logical as $UR\% = u(Z)/X$. Except, in the case of the constant function, only two parameters, noted $a$ and $b$, fully describ the uncertainty function of the method. This condensed form is very convenient when several MU estimates are calculated.

Both MU estimates, standard uncertainty and relative uncertainty, are computed from the THEOPHYLLINE dataset collected in Table 7.3. They are graphically illustrated for the standard uncertainty $u(Z)$ and for the relative uncertainty $UR\%(Z)$ in relation to the concentration $Z$ in Figure 7.8a,b, respectively. They clearly illustrate the increase of variability of MU as the theophylline concentration decreases. In this example, a power function gives a good fitness of this variation in the validated domain.

To calculate the coefficients of a power function using Excel built-in functions, such as `LINEST`, `SLOPE`, or `INTERCEPT`, the simplest method is to log-transform

**Table 7.11** Main uncertainty functions.

| Type | Standard uncertainty | Relative uncertainty | |
|---|---|---|---|
| Constant | $u(Z) = a$ | $UR\% = \left(\frac{2a}{Z}\right) \times 100$ | (7.27) |
| Proportional | $u(Z) = bZ$ | $UR\% = 2b \times 100$ | (7.28) |
| Linear | $u(Z) = a + bZ$ | $UR\% = \left(2b + \frac{2a}{Z}\right) \times 100$ | (7.29) |
| Power[a)] | $u(Z) = aZ^b$ | $UR\% = (2aZ^{b-1}) \times 100$ | (7.30) |
| | | $UR\% = (2aZ^{1+b})$ | (7.31) |

a) An explanation of these two equivalent formulas is given below.

concentration and uncertainty before the computation. This is possible with the `LN`, `LOG10`, or `LOG` functions. As shown in rows 7–9 of the Resource M worksheet, the most convenient transformation is the decimal logarithm `LOG10`. The antilog of the slope is easy to obtain, as illustrated in rows 12 and 15 of the worksheet.

**Resource M Calculation of the coefficients of a power function (Excel).**

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | **Resource M: Calculation of the coefficients of a power function** | | | | | | | |
| 2 | Theophylline (µg/l) | | | | | | | |
| 3 | *Z* | 0.05 | 0.1 | 0.5 | 1 | 2.5 | 10 | |
| 4 | *u*(*Z*) | 0.0117 | 0.0130 | 0.0350 | 0.0862 | 0.2749 | 0.5093 | |
| 5 | *U*(*Z*) | 0.0234 | 0.0260 | 0.0700 | 0.1724 | 0.5497 | 1.0186 | |
| 6 | *UR*%(*Z*) | 46.8% | 26.0% | 14.0% | 17.2% | 22.0% | 10.2% | |
| 7 | Log10(*Z*) | −1.301 | −1.000 | −0.301 | 0.000 | 0.398 | 1.000 | =LOG10(G3) |
| 8 | Log10(*u*(*Z*)) | −1.932 | −1.886 | −1.456 | −1.065 | −0.561 | −0.293 | =LOG10(G4) |
| 9 | Log10(*UR*%) | −0.330 | −0.585 | −0.854 | −0.764 | −0.658 | −0.992 | =LOG10(G6) |
| 10 | | | | | | | | |
| 11 | **Standard uncertainty function coefficients** | | | | | | | |
| 12 | Constant | 0.0907 | =10^INTERCEPT(B8:G8;B7:G7) | | | | | |
| 13 | Power | 0.7780 | =SLOPE(B8:G8;B7:G7) | | | | | |
| 14 | **Relative uncertainty function coefficients** | | | | | **Verification** | | |
| 15 | Constant | 0.1813 | =10^INTERCEPT(B9:G9;B7:G7) | | | 0.1813 | =B12*2 | |
| 16 | Power | −0.2220 | =SLOPE(B9:G9;B7:G7) | | | −0.2220 | =B13-1 | |

With the THEOPHYLLINE data (inverse-predicted by the quadratic WLS regression models), the uncertainty function coefficients are:

| Parameters | *u*(*Z*) | *UR*% |
|---|---|---|
| Constant: | 0.0907 | 0.1813 |
| Power: | 0.7780 | −0.2220 |

These values are the same as those in Figure 7.8a,b, directly obtained with Microsoft Excel by adding a trendline with the option "Power" on the graphics. When fitting the relative uncertainty power function in Excel, the *UR*% values are

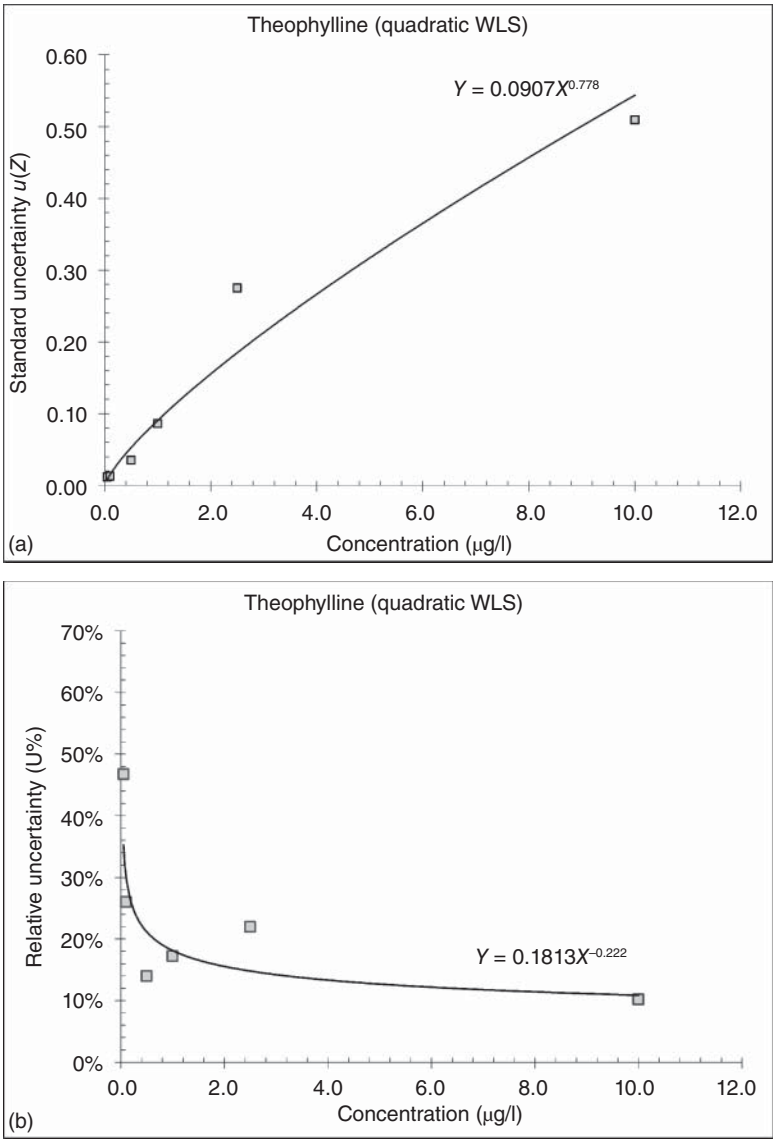



**Figure 7.8** (a) THEOPHYLLINE – standard uncertainty function, (b) THEOPHYLLINE – relative uncertainty function.

not necessarily true percentages as shown in Eqs. (7.30) and (7.31), and preliminary precautions must be taken. In Excel worksheets, the data display may not reflect the internal format. The same data worth 0.468, if the applied format is "Percentage" will be displayed as 46.8%, if it is "Number," it is 0.468. The coefficients calculated with Resource M, especially the power coefficient, do not consider the display format. If the model is used routinely, it will be necessary to ensure that the data are input in the same format as used to estimate the coefficients of the function. The

parameters of the different uncertainty functions can be connected, depending on the display format, as follows:

| | $u(Z)$ | $UR\,\%\,(Z)$ | $UR\,\%\,(Z) \times 100$ |
|---|---|---|---|
| Constant: | $a = 0.0907$ | $2a = 0.1813$ | $2a = 0.1813$ |
| Power: | $b = 0.7780$ | $b - 1 = -0.2220$ | $1 + b = 1.7780$ |

#### 7.5.2.1 How to Interpret a Power Function?

From the previous results, the power coefficient $b$ is positive regarding the standard uncertainty but negative for the relative uncertainty. While the linear function, such as the calibration straight line, is well known to analysts, the power function is less commonly considered. The latter is parametrized by two coefficients, noted $a$ and $b$. To understand their respective roles, it is necessary to recall the classic properties of power notation:

| | | |
|---|---|---|
| $x^0 = 1$ | $x^{-b} = \frac{1}{x^b}$ | $x^{-1} = \frac{1}{x}$ |
| $\sqrt[2]{x} = x^{\frac{1}{2}} = x^{0.5}$ | $\frac{1}{\sqrt[2]{x}} = x^{-\frac{1}{2}} = x^{-0.5}$ | $\frac{1}{\sqrt[3]{x}} = x^{-\frac{1}{3}} = x^{-0.333}$ |

Comparable to the Horwitz model, the power function is usually the most appropriate model to establish the relationship between MU parameters and concentration. The power coefficient $b$ for the relative uncertainty function is always $< 0$ giving a concave shape to the curve. This can be explained by the decrease of $UR\,\%\,(Z)$ when the concentration $Z$ is increasing. The curve concavity underlines this property.

Figure 7.9 illustrates various power functions when coefficient $b$ varies between [0, −0.6] and coefficient $a = 0.2$ remains constant. These values are classically encountered, for instance, if $b = -\frac{1}{2}$ it means that $UR\% = \frac{1}{\sqrt{Z}}$. The more negative $b$ is, the more pronounced the incurvature, and the more $b$ approaches 0, the more the function flattens and becomes a horizontal line, indicating that the relative uncertainty is constant and equal to $a$. The constant $a$ of the function expressed in % represents the relative uncertainty when the concentration is 1 unit. In this figure, as $a = 0.2$ the relative uncertainty is always 20% when $Z = 1$. When considering the standard uncertainty, the curve is convex. This illustrates the increase of the $u(Z)$ with concentration. The higher the coefficient $b$, the faster the increase. In practice, typical $b$ values are between 0 and 2.0. For example, $b = 0.845$ in the Horwitz model.

## 7.6 Concept of Coverage Interval

### 7.6.1 Origin of Coverage Interval

Although the initial concept of MU can be dated back to 1977, it was not until the early 2010s that a practical document was published under the reference ISO/IEC Guide 98-4 on *The Role of Measurement Uncertainty in Conformity Assessment*

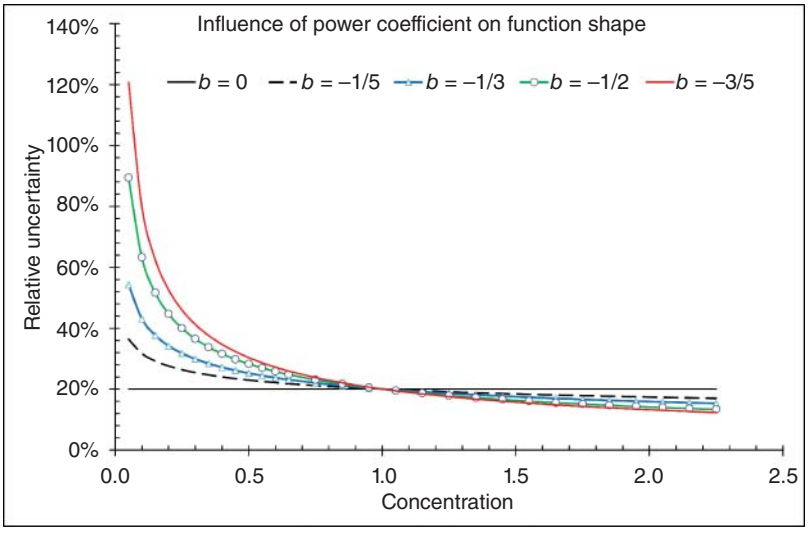



**Figure 7.9** Different power functions when power coefficient $b$ varies, and $a = 0.2$ remains constant. Concentration in arbitrary units.

to explain how MU can be used in decision-making. Two other paradigm shifts occurred during this period of slow maturation.

- The notion of a single true value for a measurand has been discarded in favor of the coverage interval.
- The definition of accuracy has been modified, making the use of this term for the MAP inconsistent with its metrological definition.

Of course, the notion of a single true value, introduced by statisticians many years ago, has practical problems. Therefore, the concept of MU implies that, for a given analyte and a given result, there is not a single true measurement value but an infinite number of possible true values scattered around the observed result. These possible true values are not equal and can be attributed to the analyte with varying degrees of reliability or probability. However, they are all compatible with the observations and with the knowledge that the analyst may have about the analyte.

This explains why the term "true value" has disappeared from GUM in the 2012 release, whereas it was present in previous versions. Instead, an analytical result is expressed as a coverage interval, defined as "the set of true values of a measurand with a specified probability based on the available information." This definition is related to the notion of statistical dispersion interval, as explained in Section 5.3.

In view of the importance of this work on MU for the industry, in 1997, the BIPM set up a specific committee to prepare and draft documents relating to the estimation of MU. This is the Joint Committee for Guides in Metrology (JCGM) which is responsible for the Guide to the Expression of Uncertainty (GUM) and the VIM. For instance, next to developing these concepts, proposals such as replacing the *TAE* with MU in expressing the results from medical laboratories have been published [31].

In practice, there is no contradiction between these two parameters. They are not mutually exclusive because they do not consider the same objects of study. The *TAE* qualifies the method's performance, while the MU seeks to qualify the measurement itself [32].

Given the ultimate goal of the analytical process, which is to provide a result that allows an informed decision, MU is therefore preferable. The use of MU as a single validation criterion for quantitative methods could be considered relevant if a universally accepted estimation procedure could emerge for the analytical sciences, which is not yet the case, as explained in Section 7.1.

MU was originally adopted for physical measurements but has only recently become familiar to chemical analysts through the ISO 17025 standard for laboratory accreditation. Uncertainty, "a parameter that characterizes the dispersion of values attributed to a measurand, based on the information used," appears more than 30 times in the 2005 version of ISO 17025 and with equal frequency in the 2017 version, while it was absent in previous versions.

Estimating MU consists in evaluating the dispersion of all the measurement values that could be obtained from a given sample. With a quick reading of this definition, one might be tempted to propose, for an estimation procedure, the interval of dispersion of measurements directly derived from the distribution function of a random variable. For example, if measurements $Z_i$ are exactly distributed according to a Normal distribution of theoretical mean $\mu$ and variance $\sigma^2$, denoted $\mathcal{N}(Z, \sigma^2)$, it is possible to confirm that 95% of the values are distributed in the following interval:

Dispersion interval

$$[\mu - 1.96 \times \sigma,\ \mu + 1.96 \times \sigma]$$

This approach is a misinterpretation of the definition of MU because such an interval assumes that $\mu$ and $\sigma$ are known and does not consider all the "information used" to obtain the measurement values. For instance, the variability of the conditions under which the measurements are made nor the sources of error, such as measurement bias, are not considered.

To circumvent this problem, it is possible to obtain a prediction interval based on estimated values of $\mu$ and $\sigma$. The prediction interval is discussed in more detail Section 5.3. Moreover, from an economic point of view, the most serious drawback is it would require systematic replication of measures to estimate coverage interval; the prediction interval estimated on one day or one sample could not be transferable to another as measuring conditions may have changed.

Given the relatively high cost of many analyses and the intense pressure on prices, laboratories would find it impractical to provide this parameter. Therefore, a more comprehensive approach to estimating MU that is transferable from one measurement to another or one sample to another is needed. To reach this goal, it must be remembered that MU is a parameter that characterizes the dispersion of measurements "once all causes of variability are taken into account." The procedure proposed by the GUM for estimating MU consists, for each measurand, in making an inventory as exhaustive as possible of all the components contributing to the

variability, as explained in Section 6.2. At this point, three prominent issues should be emphasized:

(1) The requirement in the ISO 17025 standard:

> "5.4.7.2 Testing laboratories shall also have and apply procedures for estimating measurement uncertainty, except where test methods preclude such rigorous calculations. In some cases, it is not possible to make a metrologically and statistically valid estimate of measurement uncertainty. In such cases, the laboratory should at least attempt to identify all components of uncertainty and make the best estimate possible, while ensuring that the manner of reporting does not give an exaggerated impression of accuracy".

(2) MU characterizes a measurement or a result and not a method. It contrasts with several classic validation parameters, such as the coefficient of variation of repeatability or the recovery yield, even if these parameters participate in the MU.
(3) When estimating MU, "all biases are assumed to be ignored or corrected for by appropriate and validated factors." In this case, the uncertainty of the correction factors must be included in the final evaluation, as discussed in Section 8.4.2.

In Section 7.2, a global solution for estimating MU for analytical sciences is described. It is global in that it considers all sources of uncertainty simultaneously. Such an approach is sometimes referred to as "holistic" [33]. In Section 4.1.2, the evolution of the ambiguous concept of single true value leading to the preferential use of the coverage interval to express a measurement value is explained.

Figure 7.10 shows the different notions used to define the coverage interval. The coverage factor is fixed here at 2, which implies that the *coverage interval* contains about 95% of the possible true values of the measurand.

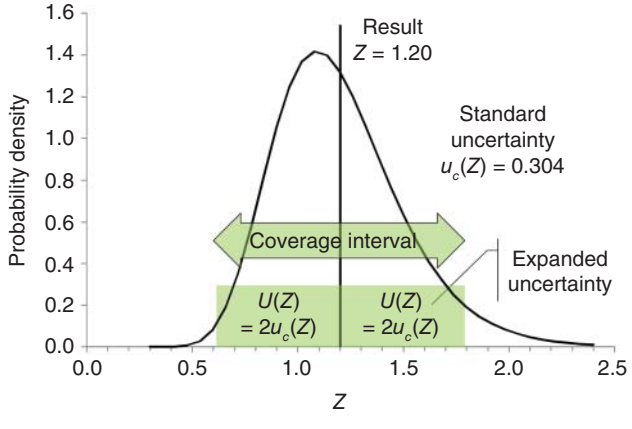



**Figure 7.10** Coverage interval and measurement uncertainty.

In this figure, which comes from simulated data, the distribution of the measurement values around the arithmetic mean is not symmetric, unlike a normal distribution. Consequently, the arithmetic mean is not at the center of the distribution on purpose and is shifted to the left. Indeed, it would be possible to compute the standard deviation, but it is not an appropriate estimate of MU. As explained before, the calculation principle of the standard uncertainty combines various standard deviations and uncertainties. The resulting statistical distribution of data is not necessarily Normal, as illustrated.

With the notations used so far, the coverage interval $[I_L; I_U]$ can be expressed in the following forms:

Diverse possible forms of the coverage interval

$$\begin{aligned} Z \pm k_{GUM} \times u(Z) \\ Z \pm U(Z) \\ Z \times (1 \pm UR\%) \end{aligned} \tag{7.32}$$

The uncertainty function can be used to calculate the coverage interval. It is satisfactory to replace the form of expression of MU with its function. At least three types of predictions can be made:

- The coverage interval of an unknown sample.
- The concentration that corresponds to a given MU.
- The concentration matches one of the bounds of coverage interval $I_L$ or $I_U$.

### 7.6.2 Coverage Interval of Given Concentration

This first type of application is illustrated by the Resource N worksheet in the case where the uncertainty function is a power function. The application to a different kind of uncertainty function can easily be adapted. According to the convention, the calculation formulas appear in columns C and F. This is an application from THEOPHYLLINE data.

**Resource N Coverage interval for a given concentration (Excel).**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Resource N: Coverage interval for a given concentration** | | | | | |
| 2 | Theophylline *UR*% function | | | | | |
| 3 | Constant | 0.1813 | | | | |
| 4 | Power | −0.222 | | | | |
| 5 | | | | | | |
| 6 | Concentration (µg/l) | *UR*%(*X*) | | Coverage interval | | |
| 7 | 0.05 | 0.352554663 | =$B$3*POWER(A7;$B$4) | 0.03237227 | 0.06762773 | =A7*(1+B7) |
| 8 | 0.25 | 0.246635236 | =$B$3*POWER(A8;$B$4) | 0.18834119 | 0.31165881 | =A8*(1+B8) |
| 9 | 1.5 | 0.165693528 | =$B$3*POWER(A9;$B$4) | 1.25145971 | 1.74854029 | =A9*(1+B9) |
| 10 | 3 | 0.14206166 | =$B$3*POWER(A10;$B$4) | 2.57381502 | 3.42618498 | =A10*(1+B10) |
| 11 | 8 | 0.114264632 | =$B$3*POWER(A11;$B$4) | 7.08588294 | 8.91411706 | =A11*(1+B11) |

### 7.6.3 Coverage Interval of Given Relative Uncertainty

This second application is illustrated by the Resource O worksheet. It also applies to an uncertainty power function because it raises some programming issues with

Excel. To predict the concentration corresponding to a given relative uncertainty, it is necessary to go through the inverse of the power function, i.e. obtain the nth root of a number. The Excel built-in function `POWER` takes `number` and `power` for input arguments and returns the power raising of `number`. For example, the `power` argument 0.5, it is possible to replace the function `SQRT`, which only returns the square root of a number. But the inverse function is not available in Excel. The simplest way to solve this problem is to use logarithms (decimal in this case), as shown in the formulation of the inverse function below.

Power function

$$UR\% = c \times Z^d$$

Inverse of the power function

$$Z = 10^{\frac{\mathrm{Log}(UR\%) - \mathrm{Log}(c)}{d}} \tag{7.33}$$

The following Resource O worksheet illustrates this calculation and how to find the concentrations and their coverage intervals for various values of relative uncertainty. This type of prediction will be used in the Section 9.2 about a possible definition of the limit of quantification (LOQ). By comparing these results with Resource N worksheet, some differences are observable:

- Inputting 0.050 µg/l in row 7 of Resource N, the relative uncertainty of 35%.
- Inputting 35%, in line 7 of Resource O, the predicted concentration is 0.0517 µg/l.

These differences are due to the rounding of the coefficients of the model and results. If all the significant figures are kept, the results are identical.

**Resource O Coverage interval for given relative uncertainty (Excel).**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Resource O: Coverage interval for given relative uncertainty** | | | | | |
| 2 | Theophylline *UR*% function | | | | | |
| 3 | Constant | 0.181 | | | | |
| 4 | Power | −0.222 | | | | |
| 5 | | | | | | |
| 6 | *UR*%(*X*) | Predicted concentration (µg/l) | | Coverage interval | | |
| 7 | 35% | 0.0517 | =10^((LOG(A7)-LOG($B$3))/$B$4) | 0.034 | 0.070 | =$B7*(1+$A7) |
| 8 | 30% | 0.10 | =10^((LOG(A8)-LOG($B$3))/$B$4) | 0.072 | 0.134 | =$B8*(1+$A8) |
| 9 | 25% | 0.24 | =10^((LOG(A9)-LOG($B$3))/$B$4) | 0.18 | 0.29 | =$B9*(1+$A9) |
| 10 | 20% | 1 | =10^((LOG(A10)-LOG($B$3))/$B$4) | 0.51 | 0.77 | =$B10*(1+$A10) |
| 11 | 10% | 15 | =10^((LOG(A11)-LOG($B$3))/$B$4) | 13 | 16 | =$B11*(1+$A11) |
| 12 | 5% | 331 | =10^((LOG(A12)-LOG($B$3))/$B$4) | 315 | 348 | =$B12*(1+$A12) |

### 7.6.4 Obtain the Limits of the Coverage Interval

In Section 8.2, this last application of the uncertainty function is meaningful when dealing with sample conformity assessment. It consists in predicting the measurement value matching one of the two bounds of a pre-defined coverage interval. For instance, the official threshold contents of a pollutant in a food product must be

below a given value. It is necessary to know the MU at this threshold to verify if a control method is adequate. Two situations must be tackled, whereas the predicted measurement can be:

- The lower bound $I_L$ of the coverage interval.
- The upper bound $I_U$ of the coverage interval.

The two formulas below deal with these two cases.

Measurement at lower bound

$$Z_{min} = \frac{I_L}{1 - UR\%} \tag{7.34}$$

Measurement at upper bound

$$Z_{max} = \frac{I_U}{1 + UR\%} \tag{7.35}$$

The relative uncertainty $UR\%(Z)$ can be computed with the uncertainty function, and the application of the formulas is direct, as the following calculations show:

| *UR* % (%) | Threshold (g/kg) | | Acceptable values must be |
|---|---|---|---|
| 40 | Maximum limit | $I_U = 5.0$ | ≤3.57 |
| 10 | | | ≤4.55 |
| 40 | Minimum limit | $I_L = 5.0$ | ≥8.33 |
| 10 | | | ≥5.56 |

Let us assume:

- The threshold value is 5 g/kg, representing either a maximum or a minimum bound.
- Two methods of analysis, one leading to 10% relative uncertainty for this content and the other 40%. The same values are used in the next chapter on sample conformity assessment and illustrated in Figure 8.4.

When using the previous formulas, it is now possible to define acceptable values when controlling a given threshold. For example, if the relative uncertainty is constant and equal to 40%, the highest acceptable concentration is 3.57 g/kg to satisfy the 5.0 g/kg threshold. This simple example with constant uncertainty is to illustrate the procedure, but the use of a power uncertainty function is not problematic.

## References


1 Standard ISO 17025:2017 (2017). *General Requirements for the Competence of Testing and Calibration Laboratories*. Genève: ISO.

2 BIPM, IEC, IFCC, ISO, IUPAC, IUPAP and OIML (2020). *Guide to the Expression of Uncertainty in Measurement (GUM)*. JCGM GUM-6 https://www.bipm.org.

3 EURACHEM-CITAC (2012). *Quantifying Uncertainty in Analytical Measurement*, 3e. https://eurachem.org/index.php/publications/guides/quam (accessed 23 July 2023).

4 Commission SFSTP, Roussel, J.-M., Botalla, S. et al. (2017). Incertitude de mesure des méthodes analytiques dans le contrôle du medicament. *STP Pharma Pratiques* 27 (4): 171–248. [in French].

5 Norme NF V03-110:2010 (2010). *Analyse des produits agricoles et alimentaires - Protocole de caractérisation en vue de la validation d'une méthode d'analyse quantitative par construction du profil d'exactitude*. Saint-Denis: AFNOR (in French).

6 EDQM-OMCL (2020). Evaluation of measurement uncertainty, annex 2: estimation of measurement uncertainty using top-down approach, annex 2.1: use of data from validation studies for the estimation of measurement uncertainty, PA/PH/OMCL, (18), 149 R1.

7 EDQM-OMCL (2020). Evaluation of measurement uncertainty, annex 2: estimation of measurement uncertainty using top-down Approach, annex 2.3: use of certified reference materials for the estimation of measurement uncertainty, PA/PH/OMCL, (18), 151 R1 CORR.

8 Standard ISO 11352:2012 (2012). *Water Quality — Estimation of Measurement Uncertainty Based on Validation and Quality Control Data*. Genève: ISO.

9 Standard ISO 13528:2015 (2015). *Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparisons*. Genève: ISO.

10 EDQM-OMCL (2020). Evaluation of measurement uncertainty, annex 2: estimation of measurement uncertainty using top-down approach, annex 2.5: use of data from proficiency testing studies for the estimation of measurement uncertainty, PA/PH/OMCL, (18), 153 R3.

11 EDQM-OMCL (2020). Evaluation of measurement uncertainty, annex 2: estimation of measurement uncertainty using top-down approach, annex 2.2: use of data from control charts for the estimation of measurement uncertainty, PA/PH/OMCL, (18), 150 R1.

12 Standard ISO 21748: 2017 (2017). *Guidance for the Use of Repeatability, Reproducibility and Trueness Estimates in Measurement Uncertainty Evaluation*. Genève: ISO.

13 EDQM-OMCL (2020). Evaluation of measurement uncertainty, annex 2: estimation of measurement uncertainty using top-down approach, annex 2.4: use of data from collaborative studies for the estimation of measurement uncertainty, PA/PH/OMCL, (18), 152 R1.

14 Da Silva, R.J.N.B., Santos, J.R., and Camões, M.F.G.F.C. (2006). A new terminology for the approaches to the quantification of the measurement uncertainty. *Accreditation and Quality Assurance* 10: 664–671.

15 Lee, J.H., Choi, J.H., Youn, J.S. et al. (2015). Comparison between bottom-up and top-down approaches in the estimation of measurement uncertainty. *Clinical Chemistry and Laboratory Medicine* 53 (7): 1025–1032.

16 H. Belmir, H. Bouchafra, A. Abbouriche, T. Saffaj, R. Ait Lhaj, M. El Karbane, B. Ihssane: Use of an uncertainty profile to validate high-performance liquid

chromatography (HPLC) for the simultaneous determination of Statins in synthetic pharmaceutical products, *Analytical Letters* 56, 2491-2504(2023) doi: https://doi.org/10.1080/00032719.2023.2177664.

**17** De Bièvre, P. (2006). Accuracy versus uncertainty. *Accreditation and Quality Assurance* 10: 645–646.

**18** Ramsey, M.H., Ellison, S.L.R., and Rostron, P. (ed.) (2019). *Eurachem/EUROLAB/CITAC/Nordtest/AMC Guide: Measurement Uncertainty Arising from Sampling: A Guide to Methods and Approaches*, 2e. Eurachem. ISBN 978-0-948926-35-8. https://eurachem.org/index.php/publications/guides/musamp (accessed 1 September 2023).

**19** Kuttatharmmakul, S., Massart, D.L., and Smeyers-Verbeke, J. (1999). Comparison of alternative measurement methods. *Analytica Chimica Acta* 391 (2): 203–225.

**20** Bugner, E. and Feinberg, M. (1992). Determination of mono-and disaccharides in foods by interlaboratory study: quantitation of bias components for liquid chromatography. *Journal of AOAC International* 75 (3): 443–464.

**21** Rozet, E., Ceccato, A., Hubert, C. et al. (2007). Analysis of recent pharmaceutical regulatory documents on analytical method validation. *Journal of Chromatography A* 1158: 111–125.

**22** Montgomery, D.C. (2009). *Introduction to Statistical Quality Control*, 6e. Wiley.

**23** Van Liedekerke, B.M., Nelis, N.J., Kint, J.A. et al. (1991). Quality control of albumin solutions by size-exclusion high-performance liquid chromatography, isoelectric focusing, and two-dimensional, immuno-electrophoresis. *Journal of Pharmaceutical Sciences* 80 (1): 11–16.

**24** Food and Drug Administration (FDA) (2018). *Bioanalytical Method Validation Guidance for Industry*. Washington, DC: Office of Communications, Division of Drug Information Center for Drug Evaluation and Research.

**25** Boyer, K.W., Horwitz, W., and R. (1985). Albert interlaboratory variability in trace element analysis. *Analytical Chemistry* 57: 454–459.

**26** Horwitz, W. and Albert, R. (2006). The Horwitz ratio (HorRat): a useful index of method performance with respect to precision. *Journal of AOAC International* 89 (4): 1095–1109.

**27** Commission of the Codex Alimentarius (CAC) (2018). Maximum residue limits (MRLs) and risk management recommendations (RMRs) for residues of veterinary drugs in foods, CX/MRL 2-2018.

**28** Thompson, M. (2000). Recent trends in interlaboratory precision at ppb and sub-ppb concentrations in relation to fitness for purpose criteria in proficiency testing. *Analyst* 125: 385–386.

**29** Codex Alimentarius (2017). Guidelines on performance criteria for methods of analysis for the determination of pesticide residues in food and feed, CXG 90-2017. Several application documents are issued. For more specific information, consult. https://www.fao.org/fao-who-codexalimentarius/codex-texts/guidelines/en/ (accessed 1 September 2023).

**30** Publications Office of the European Union (2021). Regulation (EU) 2021/808 of 22 March 2021 on the performance of analytical methods for residues of pharmacologically active substances used in food-producing animals and on the

interpretation of results as well as on the methods to be used for sampling and repealing Decisions 2002/657/EC and 98/179/EC.

**31** Oosterhuis, W., Bayat, H., Armbruster, D. et al. (2018). The use of error and uncertainty methods in the medical laboratory. *Clinical Chemistry and Laboratory Medicine* 56 (2): 209–221.

**32** Rozet, E., Marini, R.D., Ziemons, E. et al. (2011). Total error and uncertainty: friends or foes? *Trends in Analytical Chemistry* 30 (5): 797–806.

**33** Gonzàlez, A.G. and Herrado, M.A. (2007). A practical guide to analytical method validation, including measurement uncertainty and accuracy profiles. *Trends in Analytical Chemistry* 26 (3): 227–238.

# 8

# Measurement Uncertainty and Decision

## 8.1 Framework for Decision-Making

### 8.1.1 Decision *versus* Uncertainty

The word "uncertainty" has been a major source of inspiration for philosophers, psychologists, economists, and even poets. There is a long list of quotations that play with the word. It originates from the Latin verb *cernere*, which can be translated as discern, distinguish, and finally decide. The past participle of the Latin verb gave *certus* and its opposite *incertus* from which come certain and uncertain. A Latin proverb plays on these words "*amicus certus in re incerta cernitur*," a "reliable friend can be recognized in unreliable times." In its most classical sense, uncertainty means not knowing what to do. Etymologically, it is the opposite of knowing how to decide.

On the other hand, the word "decision" is derived from the Latin verb *caedere*, which means to strike, knock down, or cut. To decide is therefore to make a final cut to a question, a controversy, or a situation. Therefore, it seems paradoxical to claim that knowledge of uncertainty improves decision making.

Nevertheless, for a scientist, uncertainty is not synonymous with lack of knowledge, and in Chapters 6 and 7, we have always been careful to speak of measurement uncertainty (MU) rather than *uncertainty* in a narrow and strict sense. It is disappointing that there is such an ambiguity to name what is, in any case, an intrinsic property of a measurement value, namely the manifest dispersion of its replicates, leading to the establishment of a set of possible values for a given result.

Any measurement process has its own variability, and it is essential to accept that the analytical result is never the absolute truth, regardless of the technical improvement that has made this measurement possible. Despite the ambiguity that surrounds a single result, the latter remains fully useful, as demonstrated by the millions of analyses carried out every day, and it allows us to approach this searched-for truth. And the better the uncertainty of measurement is known, the closer the truth.

Chapters 6 and 7 have largely endorsed this affirmation, at least from a statistical point of view. As pointed out in many textbooks, deciding means facing the possibility of making a mistake and various pitfalls and dangers. For example, deciding that a lot is compliant and can be released is exposing oneself to the probability that it is, in fact, noncompliant. But a danger by itself is not the main issue, what really

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

matters is the probability of its occurrence; this is called the risk. Thus, an avalanche is a big danger with disastrous effects, but it only presents a risk in the mountains; in lowland areas the risk is zero.

Similarly, when a batch is released, if only 1 out of 10,000 units is nonconforming, the risk of receiving a bad unit is not the same as when the frequency is 1 out of 100 units. As explained in Section 6.11, the concept of risk consists of moving from the notion of frequency to that of probability in the same way it is classically achieved in risk analysis or evaluation for insurance companies.

A perfect model of decision-making would be that of the rational actor, who is a pure calculating and optimizing decision-maker. He would integrate all existing information and, on this basis, make the decision that is objectively considered the best. This model is utopian because most of its assumptions are unrealistic. Among other things, the idea that the decision-maker possesses all the information is never true in practice. Decisions are usually based on a relatively small amount of information; this is a context of limited rationality. In regard to risk analysis, there are at least two classical conceptions of uncertainty:

- The first is referred to as fundamental uncertainty and supposes that the future does not follow any predetermined laws of probability.
- The other one is called scientific uncertainty, and it has been our main subject up to now under the name of measurement uncertainty.

When considering the use of analytical data, in many situations, both definitions may coexist, especially when dealing with legal or regulatory issues. The main theories about decision-making under uncertainty come from economists, and there is a large body of literature in this area.

As far as analytical sciences are concerned, the question arose earlier in medical biology laboratories, when the practice became widespread of legally engaging the responsibility of decision-makers and prescribers suspected of having made a wrong decision. But it would be an error to believe that only medical laboratories are concerned. Today, any laboratory providing services can also be confronted with this type of liability and be prosecuted. As an example, when talking about legal liability, jurists distinguish between criminal liability, which is a source of sanctions, and civil or administrative liability, which is a source of compensation.

To make a long story short, the mission of an analytical laboratory providing services consists of carrying out analyses requested by a prescriber, who is generally responsible for interpreting the results. However, it should be noted that in certain specific areas of activity, such as forensic medicine, biomedical analysis, toxicology, and so on, the interpretation of the result is the responsibility of the laboratory itself.

A laboratory is not protected against professional negligence, carelessness, or an unforeseen event: loss of data, accidental disclosure of strategic and confidential information, failure to meet contractual deadlines, etc. In this case, the victim(s) must be compensated for the damage suffered, if it is proven. In concrete terms, this means paying damages and bearing the loss of earnings resulting from the loss of customer confidence. Civil liability insurance is designed to cover these risks and allows the laboratory to be insured against damage caused to third parties during

the laboratory's activities, to property and equipment caused by employees, or to the employees themselves.

### 8.1.2 Specification Limits and Reference Values

For industrial production lines, control of the process is essential for decision-making, and the use of analytical methods is general. However, this does not mean that all methods must be extremely efficient. The question is rather obtaining results with a level of confidence adjusted to the needs; the notion of *fitness-for-purpose* is consistent with this objective. In fact, it is just necessary to have a method that allows us to control, with a known level of risk, if the technical limit of specification or the regulatory tolerance threshold is exceeded or not.

Estimating the coverage interval of any result provides valuable information to better control the capacity of the result to permit a decision to be taken by reducing the risk of an incorrect decision. It is then easy to affirm that the uncertainty of measurement is an interesting asset for decision-making. If the example of a batch release is regarded, rational control of the risk can be performed in two steps.

- First, the decision maker uses a statistically sound sampling plan to take representative sample units from the lot to be controlled. As explained in Section 8.3, choosing a suitable plan is equivalent to calculating the probability of rejecting a nonconforming lot, i.e. efficiency of the plan. It is calculated *a priori*, by selecting a model that correctly describes the spatial distribution of the analyte in the lot (very often assumed and not measured). Generally, it is recommended to choose an efficiency of 95%, which means that the risk of accepting a not compliant lot is 5%. The idea that 5% of the released batches may be noncompliant is often misunderstood.
- Then, the decision is made on the basis of one or more measurement values obtained from the sampling units collected according to the sampling plan. The decision-maker applies the decision rule that has been chosen. For example, it consists of comparing the measurement values to a specification limit. It is at this stage that MU reduces the risk of accepting a nonconforming lot, especially if the part of the uncertainty due to sampling is included in its calculation.

Finally, the 95% coverage interval of a result is a practical way to take into account the MU.

Figure 8.1 gives a graphical illustration of this proposal. It shows that some results, which would be considered conforming if the MU is not taken into account, are in fact in the rejection interval, and vice-versa, as explained in more details in Section 8.2.2. When comparing a result to a given limit, a classical procedure consists in a significance test, often called null hypothesis testing in statistical literature.

In Section 8.2, it is demonstrated that an alternative and simpler approach to such test is just reporting the coverage interval of the result, without going through the sometimes-delicate calculation of a statistical test. Depending on the field of application there are different manners to set a threshold or limit of specification. For example, it can be:

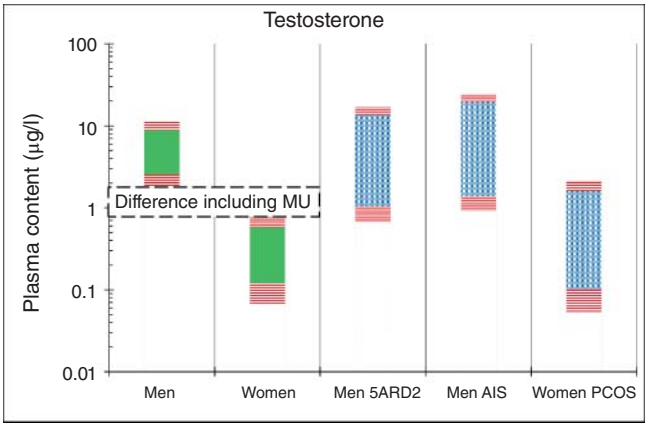



**Figure 8.1** Total circulating testosterone reference values for normal and pathological states with associated MU. Logarithmic scale.

- Technical specifications required for the proper realization of industrial objects, such as the nominal content of a drug.
- Environmental or food safety standards defined by official control authorities.
- Declarative values, such as for nutrition labelling.
- Normal and pathological limits used in medical biology.

Moreover, these limits and specifications can be unilateral or bilateral, or defined as lower or upper thresholds. To establish normal or pathological reference values in medicine, several cohorts will be recruited on a diagnostic basis: healthy patients, patients with a certain pathology, others with a second type of pathology, etc.

To obtain acceptable reference values or intervals, it is generally considered that at least 250 individuals per situation (or cohort) are appropriate. Once measurements of the biological biomarker have been collected, the reference values, whether they apply to healthy individuals or those with a pathology, are obtained by identifying the intervals that delimit 95% of the observations of each cohort. The methodology used to determine this interval is variable, and several approaches exist, either empirical or deterministic. In the two-sided case, it is no longer a reference value but a reference interval that contains 95% of the measurements, spreading from 2.5 to 97.5% percentiles.

Figure 8.1 illustrates reference intervals including [2.5%, 97.5%] quantiles for the total circulating testosterone in males and females [1]. The limits between the two groups of normal males and females (plain colors) do not overlap, as expected. Reference interval for each gender is [2.54, 8.90] µg/l for men, and [0.12, 0.58] µg/l for women, respectively. A method of analysis was developed using a LC–MS–MS technique and validated using the accuracy profile procedure [2]. It was possible to compute the relative uncertainty function that is the following power function:

$$UR\% = 0.352.X^{-0.125}$$

For each limit of the reference interval, the MU is calculated and added as a "guard band" represented with horizontal red stripes. Consequently, the differentiation zone between males and females is significantly reduced as illustrated by the dashed rectangle.

This study was undertaken to detect the use of testosterone as a doping agent for athletes. Therefore, the reference intervals for three groups of pathological individuals are added to the graphics as rectangles with crossed stripes, namely 5-α-reductase deficiency type 2 (5ARD2), androgen insensitivity syndrome (AIS) and the polycystic ovary syndrome (PCOS).

It is obvious that it is impossible to distinguish between healthy people who may have used testosterone and a certain number of pathologic individuals. The testosterone test presents a real case that illustrates classic difficulties of decision-making, either to determine the pathology of an individual, or, in this example, to control an illicit substance intake. Many reference values are one-sided and should be considered as a lower or upper limit not to be exceeded. The reference values are not the same in different countries making the decision even more difficult to make using a given biological marker.

In the area of food safety, the development of reference values, such as Maximum Residue Limits (*MRL*) established by the European Commission [3] and illustrated in Section 9.3 is somewhat more complex because it requires the combination of three sources of data:

- Toxicological reference values (TRV). These are toxicological thresholds established by international or national authorities such as the World Health Organization (WHO) that reflect a dose–response relationship or a point calculated from such a relationship. In the dietary risk analysis, TRVs are specifically estimated for oral ingestion in water and food, and most often for low level exposure over extended periods, such as lifetime. Most classical naming of the TRV is acceptable daily intake (ADI) or tolerable daily intake (TDI) when considering intentionally added substances, such as additives.
- Food composition data for nutrients and contaminants. There are many databases available online. However, there are two problems related to the collection of this type of data. On the one hand, the very incomplete state of knowledge on the composition of foods, since a food as *simple* as wine contains nearly 250 chemical constituents, including alcohol and water. On the other hand, the notion of "average" composition is very misleading since data significantly vary and are not distributed according to a normal law. This last remark is even more exact for contaminants and pollutants.
- Food consumption data. It is probably the most difficult information to obtain because the collection cost is high and the representativity poorly guaranteed. There are multiple strategies for investigating individual or family diet over varying periods of time. They all present biases, and it is by crossing different surveys that it is possible to reach some accuracy.

The three types of data are combined to define exposure levels. As the data are not normally distributed, such studies raise complex statistical issues which have been

previously addressed by one of us [4]. Commonly applied methods allow to identify the most probable sources of risk, i.e. main foods which are the key-sources of risk. It is then up to the regulator to decide which specification limit should apply for which analyte in which food.

It is a rather complex process based on much data, which explains why sanitary limits can often be modified. Another body in charge of this mission is the Codex Alimentarius, which also publishes internationally accepted *MRLs* [5]. Local regulation may also put forward other thresholds. But if the country is a member of the WTO, these local thresholds must be justified and not appear as technical barriers to trade (TBT).

Like the European Union, Codex also defines performance requirements for the methods that will be used to enforce the established specification limits. This task is specifically assigned to the Codex Committee on Methods of Analysis and Sampling or CCMAS. For example, for a method to be acceptable it must fulfill following requirements:

| If the *MRL* is, | then the LOQ must be |
|---|---|
| ≥0.1 mg/kg | $\leq \frac{MRL}{5}$ |
| <0.1 mg/kg | $\leq \frac{2 \times MRL}{5}$ |

The Codex Alimentarius approach is missing a guideline on the procedure to be applied by laboratory in demonstrating that the proposed method meets required criteria and properly proves its ability to perform an inspection.

### 8.1.3 Role of the Analytical Report

The analytical report is another essential element in the relationship between the laboratory and its client facing a decision. It is a contractual document and therefore has a formal aspect complementary to the technical work of the analyst. The content is described in documents relating to accreditation or quality control.

For example, in the context of the accreditation of medical laboratories, ISO 15189 standard introduces the concepts of pre-analytical, analytical, and post-analytical procedures. The reception of an analysis request sent by a prescriber or the sampling form is the pre-analytical part. The analysis report or report of the result is the post-analytical part. As addressed in the ISO 17025 and 15189 standards, elaborating the analytical report appears to be a formal administrative task. In fact, it involves the civil liability of the laboratory.

To ascertain this responsibility, a set of statements must appear in the report, like the regulatory name of the analyte, the unit used, etc. This is perfectly detailed in the regulatory texts, standards, or guides provided by the accreditation bodies. Inspection bodies are very vigilant about this documentary and/or regulatory aspect. The requirements may vary according to the accreditation program. In general, it is required that the expression of results be unequivocal.

For example, the method of analysis used must be mentioned, and even reagents, whenever they may affect the expression of the result, as well as when required by regulations. Many statements seem odd, such as: "For quantitative results, where appropriate, the analytical performance of the method may be stated." Does this mean that for nonquantitative methods, such as identification assays or diagnostic tests, the mention of sensitivity and specificity is superfluous? The analysis report is printed on a laboratory-headed paper, including the regulatory references, and is signed.

It can only be communicated after the procedure called *analytical validation*, which is different from the *method validation* described in Chapter 5. Analytical validation is carried out under the responsibility of the head of the analytical service. It consists in checking if the conditions of execution of the analysis conform with preestablished specifications. It can only be carried out after checking the indicators of instrument qualification and the results of the internal quality control. A complementary *scientific* validation can also be carried out under the responsibility of a accreditated specialist in most of the laboratories. This involves ensuring the consistency of the results of all the analyses performed.

Up to now, the indication of the MU in a significant part of the analytical report is not required. While ISO 15189 does not mention it, ISO 17025 (clause 7.8.3) states that:

> "where appropriate, the measurement uncertainty expressed in the same unit as the measurand or in a term related to the measurand (e.g. as a percentage), where it is significant for the validity or application of the test results, where required by customer instructions, or where the uncertainty affects conformity with the limits of a specification."

Similar remarks can be found in the documents written by various accreditation authorities.

## 8.2 Sample Conformity Assessment

### 8.2.1 Define the Decision Rule

Conformity assessment refers to the "activity to determine whether specified requirements relating to a product, process, system, person or body are filled." The issue of conformity assessment is formally addressed in the standard ISO 17025:2017 by means of a clearly stated decision rule defined as a (clause 3.7) "rule describing how measurement uncertainty is accounted when stating conformity with a specific requirement."

The same standard describes the context of the application of this rule (clause 7.1.3) "when the customer requests a statement of conformity … the decision rule must be communicated to and agreed with the customer, unless inherent in the requested specification of standard." Moreover, (clause 7.8.6.1) "when a statement

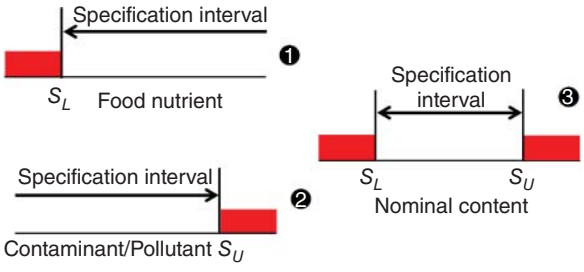



**Figure 8.2** Three ways to assess sample conformity to a unilateral or bilateral specification interval.

of conformity … is provided, the laboratory must document the decision rule it employs, taking into account the level of risk associated with the decision rule, and apply the decision rule" but no further explanation is given on how to practically apply the decision rule.

One of the most frequent examples of a decision is the verification of manufactured lot conformity. In other words, the demonstration that the concentration of a target analyte measured in a sample (that may contain several pooled sampling units) complies with the value of a limit of specification, i.e. a threshold. According to the situation, it can be a lower specification limit, denoted $S_L$, an upper specification limit $S_U$, or, more rarely, a specification interval $[S_L, S_U]$; the first two define a unilateral interval, the last a bilateral interval.

Figure 8.2 illustrates these three types of intervals. For example, for the control of pollutants or contaminants, the concentration must remain below an upper limit $S_U$, which classically the legislator has defined, such as the European regulation 470/2009 "laying down Community procedures for the establishment of residue limits of pharmacologically active substances in foodstuffs of animal origin" [6].

Compliance is therefore declared in relation to the specification interval. Usually, the decision rule is more complicated than simply looking at one reading. It is often established in relation to the sampling plan, and several measurements may be used. The sampling plan was previously established to compute what is called the efficiency of the plan, i.e. the probability of detecting a nonconforming lot.

This efficiency may vary depending on the number of accepted defects. For instance, it can be a zero-defect plan or where a certain number of results nonconforming to the specification interval are acceptable. It is not the purpose here to discuss further the theory of sampling plans, as many books and several series of guidelines describing decision rules for measurement or attribute-based control plans for different decisional contexts, such as industrial production of clinical biology, exist.

In many laboratories, a common procedure is to proceed in steps. First, an initial measurement is made. When the latter is not within the acceptability interval, so that the sample should is declared nonconforming, it is accepted to obtain a confirmatory replicate. If the first measurement is invalidated, a third one will be done.

From there, the decision rule can be fundamentally modified (and not always reported). This iterative procedure raises various problems, including statistical, that are rarely addressed. For example, when the decision rule is based on a single result, it is a direct violation of the rule. Or, when the result is given as an

average of replicates, the calculation of the associated MU must involve an adjusted intermediate precision standard deviation because these replicates are generally obtained in intermediate precision conditions. A discussion on the consequences of measurement replication on uncertainty is presented in Section 8.4.3.

### 8.2.2 Guard Band Concept

In 2012, the BIPM published a document known as JCGM-106 on the role of MU in conformity assessment [7]. Most of the concepts developed here have been taken up in the general ISO guide 98-4:2012 or the EURACHEM guide specifically dedicated to using MU for compliance assessment using chemical measurements [8, 9].

An important concept introduced in the this JCGM document is the "guard band" that corresponds to the expanded uncertainty estimated for the concentration at the limit of specification. It is added or subtracted to the limit of specification depending on whether it is a lower or upper specification limit.

Figure 8.3 illustrates this concept and associated vocabulary in the case where an upper limit of specification must be checked, for example, for pollution control. To run a control, the main issue is calculating the guard band. In the example of Figure 8.3, it is equal to the expanded uncertainty of the specification limit, which should cover 95% of future measurements around this limit. If the uncertainty function of the method is available, this calculation is easy. This approach seems simple but is not without problems in its application to the analytical sciences because the uncertainty varies rapidly with the concentration. Even for a small range of concentrations, MU can differ significantly. This property of the MU can have important consequences for the guard band calculation, as illustrated by the following simulated data.

Let's assume two different compliance objectives. In the case illustrated on the left, Figure 8.4, the objective is an upper limit that should not be exceeded. The decision rule is simple: if the analysis reveals that the concentration is too high, the batch is rejected. On the right side of Figure 8.4, it is a minimum content that must be reached, and the decision rule is reversed. In both cases, the limit of the specification is 5.00 in arbitrary units. The interpretation of this figure is as follows:

- The specification limit appears as a thick vertical line.
- The limit of the guard band is marked by a dashed vertical line.

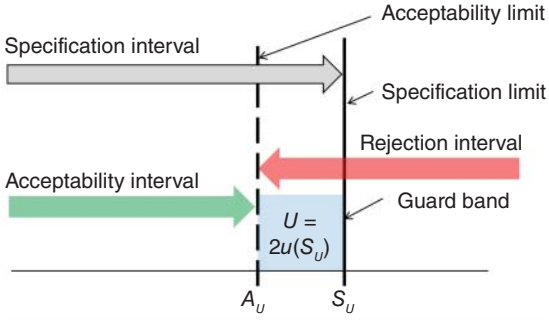



**Figure 8.3** The guard band concept introduced by the JCGM.

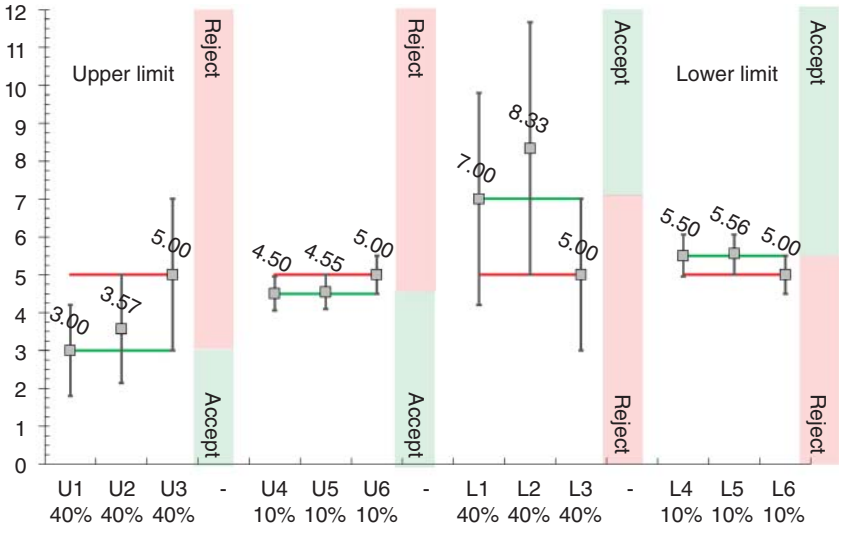



**Figure 8.4** Influence of MU on acceptability or rejection intervals. The coverage interval of each measurement value is shown as vertical error bars. Three different situations are represented for each case.

- The size of the guard bands (blue square) depends on the expanded uncertainty.
- The rejection interval is symbolized by a light red rectangle starting from the largest guard band. The acceptability interval is a light green rectangle.

In Figure 8.4, the simulated results of two control methods are illustrated, a well-performing one where the relative uncertainty is $UR\,\% = 10\%$ and a less efficient one with $UR\,\% = 40\%$. These percentages are reported below the 12 simulated examples of the figure. Depending on whether the relative uncertainty is 40 or 10%, the guard band is equal to 2.00 or 0.50. It is added or subtracted depending on whether it is a lower or upper specification limit. Corresponding examples are labeled U for the "Upper" limit and L for "Lower."

From there, three situations are considered.

- Cases U1 and U4, L1 and L4. They correspond to the decision rule proposed by the BIPM: the measured content of a sample must be less or equal to the limit of the rejection interval to be in conformity. If the uncertainty is 40% the rejection limit is $5.00 - 2.00 = 3.00$. Thus, case U1 of Figure 8.4 will be declared compliant since its content is precisely 3.00. But when its coverage interval is calculated with the same constant uncertainty of 40% – which is not always the case – it is between [1.80, 4.20]. The upper limit is, therefore, below the specification limit. This sample is penalized.
- This means that a sample presenting a slightly higher value than U1 could give a result less than or equal to 5.00 and therefore be compliant. Furthermore, it is possible to predict what sample can have the upper limit of its coverage interval reaching exactly 5.00.

- Cases U2 and U5 were simulated using this calculation, as described in Section 7.6.4. The reasoning is reversed for cases L2 and L5 in Figure 8.4.
- This effect of relative uncertainty is much less pronounced when it is only 10%.
- Finally, cases U3 and U6 and L3 and L6 show that if the result is equal to the specification limit, 100% of the values included in the coverage interval are in the rejection zone.

According to these diagrams, the four samples numbered 2 and 5 are the most problematic. These samples should be declared compliant since the limit (upper or lower) of their coverage interval corresponds to the specification limit. In other words, only 2.5% of the measurements that could be obtained would not respect the objective. It seems that the principle of the guard band predicted from the expanded uncertainty of the limit of specification applies rather poorly to the analytical sciences for two reasons:

- The uncertainty can be quite high, mainly when close to the LOQ, while regulators try to enforce limits in the vicinity of the LOQ.
- The MU can significantly vary even over a small concentration interval.

It would be more interesting to define the guard band by calculating, from the uncertainty function, at which value the lower or upper bound of the coverage interval is equal to the specification limit, as explained in Section 7.6.4.

This remark underlines the interest of being able to determine beforehand the uncertainty function. In any case, the ISO guide 98-4, which addresses this topic, is mainly oriented toward industrial production, where MU is usually reduced and constant. Indeed, it proposes integrating the statistical characteristics of the production process or other phenomena of interest.

For that, this information must be known *a priori* and described by a probability law, normal, uniform, or other. To account for the variability of the manufacturing process, the guide recommends applying a Bayesian approach which allows integration in the same model of the variability of the production and that of the measurement. But for a service laboratory that receives samples of unknown origin from very varied contexts, the notion of variability due to the manufacturing process is inoperative.

The role of MU in decision-making cannot be discussed without pointing out the reluctance of prescribers to use it. However, with the generalization of the coverage interval, the end-user will be able to benefit from this additional information in a decision-making process or to better interpret a laboratory result. Because the coverage interval contains a probability, it is an indication of the level of confidence a decision maker has to implement a decision rule.

For example, in the context of official food control, the European Commission published a regulation in 2021 that describes how to use MU to enforce *MRL* legislation [10]. As explained in Section 9.3, it results in a new way of calculating two criteria, namely $CC\alpha$ and $CC\beta$, now based on MU that will replace the method proposed in 2002.

## 8.3 Sampling Uncertainty

### 8.3.1 Sampling and Heterogeneity

Sampling is a specific field of statistics. There are many textbooks or standards addressing the selection and organization of sampling plans from various technical or industrial perspectives. It is also suggested to refer to standards, regulatory texts, or professional guides or recommendations for defining the best-adapted sampling plan for a given situation. As already pointed out in Chapter 7, a semantic problem should be clarified about the term "sample" widely used in analytical sciences:

- In the laboratory, the sample is the entity or the object on which an analysis is performed. Unfortunately, it may refer to the proper object sent by a user as well as the result of the preparation performed on this object before introduction into the measuring device, such as dilution, solvent extraction, and mineralization.
- In sampling theory, the sample is a set of sampling units taken out of a batch or a population according to some specific rules.

Figure 8.5 illustrates the various terms used in sampling theory to elaborate a sampling plan. Finally, in the customary discussion, the term sample can be used instead of sampling unit, test portion, or the result of its preparation. On the other hand, the term sampling can refer to the operation of collecting a sample or the result of the operation.

To avoid ambiguity, the term sample is used in this chapter to refer to the laboratory meaning of "working sample". According to sampling theory, the number of sampling units depends on the mathematical model chosen to describe the representativeness of the analyte in the situation under study.

In analytical sciences, it is mainly the spatial distribution of the analyte that matters. The general procedure is to produce a statistical model that best describes this spatial distribution and derive the efficiency curve of the sampling plan, i.e. the curve that gives the probability of identifying a nonconforming batch. A sampling plan is

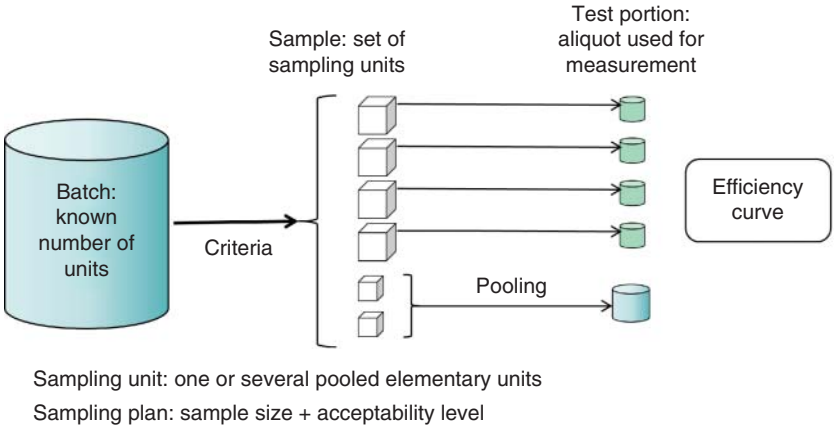



**Figure 8.5** Basic sampling vocabulary.

considered efficient when the probability is equal to or above 95%. Many laboratories are not responsible for sampling. Their role is limited to the analytical part of the analysis process, and sampling uncertainty is not under their responsibility. However, for some others, such as clinical laboratories, this is not the case.

When sampling uncertainty is required, the spatial distribution of the analyte is the main source of uncertainty to be accounted for. For instance, sampling uncertainty may be required for the official sanitary control of foods. The spatial distribution relates to homogeneity defined by IUPAC, i.e. "the degree of regularity with which a property or analyte is distributed in a quantity of the material being analyzed." Except for unusual cases, the spatial distribution of the analyte is not regular, either at the lot or sample level and even in liquid products.

Since the 1970s, four types of spatial distribution have been identified and illustrated in Figure 8.6. The so-called uniform distribution is mostly imaginary and does not exist, even after homogenization. This is unfortunate because it is the only case where the sampling uncertainty can be estimated as zero. If the most aggregated distribution is called contagious, it is because it is used as a model in epidemiology to describe the contamination clusters. These graphics show that homogeneity (or heterogeneity) depends on two parameters:

- Spatial distribution of the analyte.
- The ability of the analyte to form aggregates.

Ideally, prior knowledge of these two properties would help optimize the sampling strategy. A theory on sampling uncertainty was postulated by geochemists who introduced the concept of a sampling constant to characterize the spatial distribution of an analyte [11].

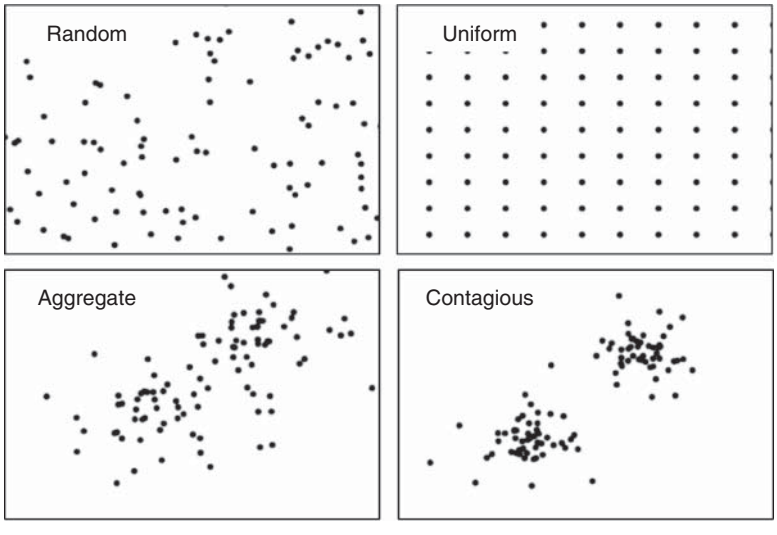



**Figure 8.6** Main types of spatial distribution of an analyte in a batch or population.

Many papers were published on this topic because the profitability of the mining industry largely depends on the laboratory's ability to detect nuggets and the beginning or end of a vein. Though the motivation was economic, it paved the way for an essential reflection on sampling.

The proposed model assumes that a lot is formed by a population of $N$ fragments with $1 \leq n \leq N$. The analyte concentration in each fragment is $Z_n$ and contributes to the heterogeneity of the lot by its deviation from the global mean $\overline{Z}$. Each fragment also has its mass $m_n$ and these parameters are related by the following equations:

The total mass of the lot

$$M = \sum_N m_n \tag{8.1}$$

Individual contribution to heterogeneity

$$H_n = \frac{Z_n - \overline{Z}}{\overline{Z}} \times \frac{N \times m_n}{M} \tag{8.2}$$

Global mean

$$\overline{Z} = \frac{\sum_N (m_n \times Z_n)}{M} \tag{8.3}$$

Weighting coefficient

$$W_n = \frac{N \times m_n}{M} \tag{8.4}$$

Weighted variance

$$s_H^2 = \sum_N \frac{(Z_n - \overline{Z})^2 W_n}{\overline{Z}^2 \sum_N W_n} \tag{8.5}$$

These parameters are interpreted as follows:

- When the analyte does not form aggregates, the mean of the $H_n$ is zero, and the variance $s_H^2$ is used to define a constitutional heterogeneity coefficient $HC$:

$$HC = s_H^2$$

- When the analyte forms aggregates, the variance $s_H^2$ corresponds to a distributional heterogeneity ($HD$), which is related to the constitutional heterogeneity by the following relation, where $FR$ is a positive aggregation coefficient.

$$HD = HC \times \frac{1 + FR \times FS}{1 + FR}$$

If $FR = 0$, there is no aggregation, and $HD = HC$.

Parameter $FS$ is a segregation factor, with $1 \geq FS \geq 0$. If $FS = 1$, the segregation is complete, and $HD = HC$. In all other cases, $HD < HC$ and material mixing or homogenizing can reduce its value. From these parameters, a sampling constant can be defined. This is a very elaborate theory mainly inspired by geostatistics, a special field of statistics [12].

The derivation of sampling uncertainty from this model is also complex. It is easily applicable to particulate solids but presents important difficulties for analyzing

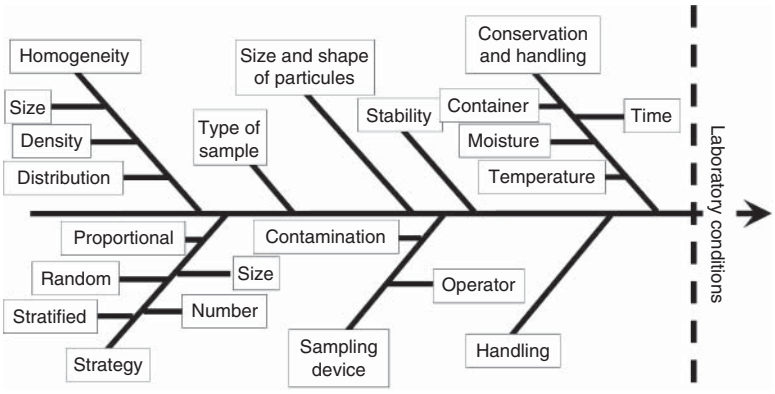



**Figure 8.7** Cause to effect diagram of the sampling operation.

living materials, such as foods or biological tissues and fluids. In this case, the underlying assumptions concerning the sampling constant or the mode of aggregation of the analyte are not easy to verify or even fulfilled. It involves a significant analytical cost that is not always justified [11].

According to the general GUM procedure for MU estimation described in Section 6.2, the first step consists of listing the different sources of uncertainty specific to the sampling and constructing a cause-to-effect diagram, like Figure 7.2. The diagram adapted to sampling is reproduced in Figure 8.7. It implies that studying each component would require a daunting number of measurements.

According to the GUM general procedure, the sources of uncertainty can be grouped. In Section 7.2, it is stated that global analytical MU can generally be expressed as the combination of three main standard uncertainties, namely:

| | |
|---|---|
| The analytical process (analytical uncertainty) | $u(Z_m)$ |
| The sampling process (sampling uncertainty) | $u(Z_s)$ |
| The measurand definition (definitional uncertainty) | $u(Z_d)$ |

From this statement, the combined standard variance of any analytical measurement can be described by Equation (8.6):

Combined standard variance

$$u_c^2(Z) = u^2(Z_m) + u^2(Z_s) + u^2(Z_d) \tag{8.6}$$

Leaving aside the definitional uncertainty, which depends very much on the traceability of the measurand to the SI unit system as explained in Section 4.1.3, the standard variance of a result can be written in a simplified form:

The simplified form of the standard variance

$$u^2(Z) = u^2(Z_m) + u^2(Z_s) \tag{8.7}$$

If the part due to the analytical process $u(Z_m)$ has been estimated beforehand, which is the case when the MAP and uncertainty function has already been established, it is easy to extract the sampling uncertainty:

$$u^2(Z_s) = u^2(Z) - u^2(Z_m)$$

$$u(Z_s) = \sqrt{u^2(Z) - u^2(Z_m)}$$

Recently, several procedures adapted to the analytical sciences have been published. These propositions are diverse and often require sophisticated means to collect the data and perform the calculation [8, 13].

### 8.3.2 Procedure of Homogeneity Check

In the assessment of sampling uncertainty, the estimation of homogeneity is one of the key issues. Unfortunately, the deterministic modeling described above is not universal, but another can be proposed. It is rather easy as it is based on statistical tools already developed in Chapters 3 and 5, such as ANOVA and tolerance intervals. The starting point is Annex B of the ISO 13528 standard which describes an experimental design for evaluating the homogeneity of the test materials used in proficiency testing schemes (PTS).

This Annex is entitled "Verification of homogeneity and stability of proficiency test entities." It describes an operating procedure for the verification of homogeneity that can be adapted as follows:

- From the lot used for the study, select $N$ samples, i.e. preferably called sampling units, with $N \geq 10$.
- For each sample, prepare two test portions using appropriate techniques to minimize differences between them.
- Analyze the $2 \times N$ test portions in random order but under repeatability conditions.
- Applying a one-way random effect ANOVA, compute the grand mean $\overline{\overline{Z}}$, within-sample standard deviation $s_r$ and between-samples standard deviation $s_B$ as described in Section 3.2.

To decide if the homogeneity of the lot is acceptable, the standard ISO 13528 proposes to compare the inter-sample standard deviation $s_B$ to a reference standard deviation $\sigma$ called "standard deviation for proficiency assessment," which is a "measure of dispersion used in the evaluation of results of proficiency testing, based on the available information" and is supposed to allow an assessment of laboratory proficiency. The acceptance criterion used to decide if the lot is sufficiently homogeneous is:

Acceptance criterion

$$\frac{s_B}{\sigma} \leq 0.3$$

Squared criterion

$$\frac{s_B^2}{\sigma^2} \leq 0.1$$

The squared acceptance criterion can be interpreted as follows: the between-samples variance must not exceed 10% of the reference variance. In practice, the proposed acceptance criterion seemed rather strict, and in a more recent standard applied to microbiology under revision (ISO 22117), it was set at 0.5 so that the between-samples variance must not exceed 25% of the reference variance. To adapt this procedure to sampling uncertainty, a simple suggestion is to replace this ratio with the variance ratio $A$ already described (Eq. 5.5, page 126). The acceptance criterion becomes:

Possible acceptance criteria

$$A = \frac{s_B^2}{s_r^2} \leq 0.1 \text{ or } 0.25 \tag{8.8}$$

In numerous accuracy profile examples shown in Chapters 5 and 7, the variance ratio $A$ is much higher than one of these acceptance limits. However, these data are collected on purpose under intermediate precision conditions and not repeatability. Therefore, it is mandatory that the measurements made to check the homogeneity be obtained under repeatability conditions to avoid any additional source of variation being interpreted as heterogeneity. The standard variance $u^2(Z)$ is described by Equation (8.7), and the decision whether it is necessary to incorporate the sampling uncertainty into combined uncertainty can be made on this acceptance criterion. If the $A$ ratio is higher than these limits of 0.1 or 0.25, sampling uncertainty is significant. The second proposal is to calculate the parameters of the β-expectation tolerance interval (β-ETI) from the data collected according to the experimental design described in Annex B of ISO 13528 standard, as explained in Section 7.2. The standard deviation of this interval denoted $s_{IT}$ can be directly interpreted as an estimate of the sampling uncertainty $u(Z_s)$.

### 8.3.3 Example of Copper in Wheat Flour

Within the framework of the official control of pesticides in wheat flour, the analysis of copper was used to check the homogeneity of a T65 wheat flour lot. Twenty samples were taken from the batch of wheat flour at different randomly selected locations for this check. No prior homogenization was applied to be consistent with the future conditions of use. From each sample, two test portions were analyzed by flame atomic absorption spectrophotometry. This method was selected as it is fast and inexpensive with a well-established analytical uncertainty. Table 8.1 is a compilation of 40 measures performed on separate test portions issued from 20 control samples denoted S01 to S20.

The data are input into the Resource H worksheet after increasing the number of series from 6 to 10. The best way is to add rows in the middle of an existing range so that all formulas are automatically updated, except in columns `D` and `E`, where formulas must be manually copied. From the computed parameters, several conclusions can be drawn. Table 8.2 shows the β-ETI results using the notations of this book. When verifying the acceptance criteria proposed by formula (8.8), it is possible to conclude that the batch is heterogeneous because the variance ratio $A$ is much above 0.1 or 0.25. Sample homogeneity needs to be evaluated.

**Table 8.1** COPPER – measurements in T65 wheat flour (mg/kg)[a].

| Control | Measure 1 | Measure 2 | Control | Measure 1 | Measure 2 |
|---|---|---|---|---|---|
| S01 | 4.2 | 4.3 | S11 | 4.3 | 4.3 |
| S02 | 4.5 | 4.3 | S12 | 4.2 | 4.2 |
| S03 | 4.1 | 4.5 | S13 | 4.5 | 4.3 |
| S04 | 4.1 | 4.2 | S14 | 4.3 | 4.5 |
| S05 | 4.1 | 4.3 | S15 | 4.2 | 4.4 |
| S06 | 4.2 | 4.2 | S16 | 4.2 | 4.3 |
| S07 | 4.3 | 4.1 | S17 | 4.2 | 4.1 |
| S08 | 4.1 | 4.5 | S18 | 4.1 | 4.3 |
| S09 | 4.6 | 4.7 | S19 | 4.5 | 4.5 |
| S10 | 4.5 | 4.6 | S20 | 4.6 | 4.5 |

a) Unpublished personal data.

**Table 8.2** COPPER – statistical homogeneity parameters.

| Parameters | Symbol | Value | Standard deviation | Variance |
|---|---|---|---|---|
| Average content (mg/kg) | $\overline{\overline{Z}}$ | 4.316 | | |
| Within sample std. dev. | $s_r$ | | 0.1291 | 0.0167 |
| Between-samples std. dev. | $s_B$ | | 0.1175 | 0.0138 |
| Intermediate precision std. dev. | $s_{IP}$ | | 0.1746 | 0.0305 |
| β-ETI standard deviation | $s_{TI}$ | | 0.1777 | 0.0316 |
| Effective number of measures | $N_E$ | 31.7 | | |
| Variance ratio | $A$ | 0.828 | | |

Figure 8.8 illustrates this heterogeneity. At this level of concentration, the relative uncertainty of the copper determination in wheat flour is 3.2% giving a coverage interval of [4.18, 4.15] mg/kg. This interval is shown in Figure 8.8 as two dashed lines around the average marked by a thick solid line and underlines the heterogeneity.

It was already demonstrated in Section 7.2.4 and Equation (7.5) and following, that the standard variance of the β-ETI $s_{TI}^2$ can be regarded as a proper estimation of standard variance $u^2(Z)$. This property is used to give an estimate of the left part of Equation (8.7). On the other hand, the uncertainty function obtained from the copper data gives 3.2% for the relative uncertainty of a measurement of 4.31 mg/kg. This value will allow us to calculate the analytical standard uncertainty $u^2(Z_m)$ without forgetting the coverage factor of 2. The whole calculation is summarized below:

| Source | Symbol | Value | Formula |
|---|---|---|---|
| Average content | $\overline{\overline{Z}}$ | 4.316 | |
| Analytical process | $UR\%(Z_m)$ | 3.20% | $0.1069 = \frac{(4.316 \times 0.032)}{2}$ |
| | $u(Z_m)$ | 0.06906 | |
| | $u^2(Z_m)$ | 0.00477 | $0.06906 \times 0.06906$ |
| Simplified model | $u^2(Z) = s^2_{TI}$ | 0.03158 | Equation (8.7) |
| Sampling | $u^2(Z_s)$ | 0.02681 | $u^2(Z_s) = u^2(Z) - u^2(Z_m)$ |
| | $u(Z_s)$ | 0.16373 | $u(Z_s) = \sqrt{u^2(Z) - u^2(Z_m)}$ |
| | $UR\%(Z_s)$ | 7.59% | $UR\%(Z_s) = 0.164/4.316$ |

In conclusion, for this particular case, the sampling uncertainty represents a significant contribution to the total combined uncertainty. Its relative contribution is about 7.59%, and is higher than the relative uncertainty due to the analytical process, i.e. 3.20%. The models developed for geochemical measurements, and briefly presented at the beginning of this chapter, introduced the concept that sampling uncertainty depends on a sampling constant which is an invariant, independent of the concentration level. This would mean that the 7.59% value can be used for any copper content but also analytes other than copper. However, this statement needs to be verified.

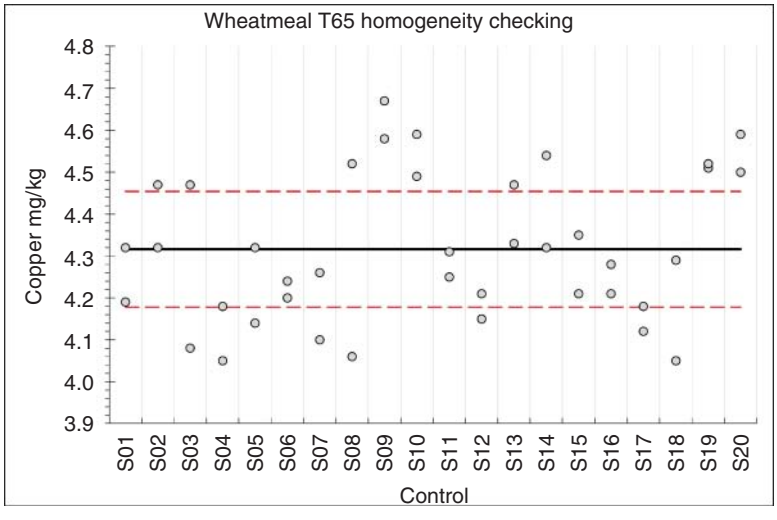
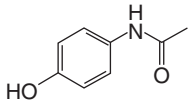



**Figure 8.8** COPPER – distribution of measurements and control number.

## 8.4 Measurement Uncertainty: Special Issues

### 8.4.1 Influence of the Calibration Model




Paracetamol

The use of a calibration model to predict concentrations seems perfectly normal and expected for analysts. However, the question remains whether the chosen calibration model can be a significant source of uncertainty. A publication on the validation of a paracetamol method for drug control may help to answer this question [14]. The following table summarizes the characteristics of the PARACETAMOL dataset.

In the original publication [14], it is not stipulated if a replicate consists of applying the entire analytical operating procedure on unique test portions or whether replicates are limited to repeated measures on a single, fully prepared test portion. We will consider a replicate a "full replicate" as defined in Section 8.4.3. Complete dataset description is reported in Table 8.3.

Using the calibration experimental data, it is possible to fit two types of models. The question of selecting the best calibration model has already been explained in Section 8.1 and resolved using the MAP procedure. When $X$ is the paracetamol content of the calibration solutions expressed in mg/ml, and $Y$ is the instrumental response expressed in absorbance units, two models can be fitted:

Linear curve

$$Y = a_0 + a_1 X + e$$

**Table 8.3** PARACETAMOL – description of the dataset.

| Title | PARACETAMOL |
|---|---|
| Reference | [14] |
| Measurer | Paracetamol concentration in a drug, expressed in mg/ml. |
| Method | UHPLC (ultrahigh-performance liquid chromatography) coupled with UV detection |
| Validation area | Between 25 and 150% of the nominal value of the finished product of 3.25 mg/ml (0.8125 to 4.875 mg/ml) |
| Acceptance interval | ±5%. Value often used for medicine quality control |
| Validation materials | 4 synthetic validation materials containing 0.8125, 1.95, 3.25 and 4.875 mg/ml respectively prepared by weighting. |
| Validation design | Series ($I = 3$), replicates/series ($J = 5$), levels ($K = 4$) |
| Calibration design | Series ($I' = 3$), replicates/series ($J' = 3$), levels ($K' = 6$) |
| Total number of measures | 54 measurements on calibration solutions 60 measurements on materials with known contents. |

**Table 8.4** PARACETAMOL – calibration model coefficients.

| Series | Model | $a_0$ | $a_1$ | $a_2$ | $r2$ |
|---|---|---|---|---|---|
| Series 1 | | 16.984 | 52.843 | | 0.996 |
| Series 2 | Linear | 16.562 | 53.075 | | 0.996 |
| Series 3 | | 16.205 | 53.319 | | 0.995 |
| Series 1 | | 0.712 | 67.862 | −2.641 | 0.997 |
| Series 2 | Quadratic | 1.550 | 66.932 | −2.436 | 0.999 |
| Series 3 | | 0.826 | 67.514 | −2.496 | 0.988 |

Quadratic curve

$$Y = a_0 + a_1 X + a_2 X^2 + e$$

Table 8.4 combines the coefficients for each serial (or daily) model. In this example, coefficients are stable from one day to another. The opposite conclusion was made with the THEOPHYLLINE study. However, selecting the best calibration model is difficult just by looking at the coefficients of determination and is even impossible in this case.

The MAPs obtained using the two models are plotted in Figure 8.9. Once more, the decision is easy as it is obvious that the linear model is unsatisfactory while the quadratic calibration model should be applied. The conclusion is evident, the method accuracy profile can be successfully used to achieve the final optimization of the method.

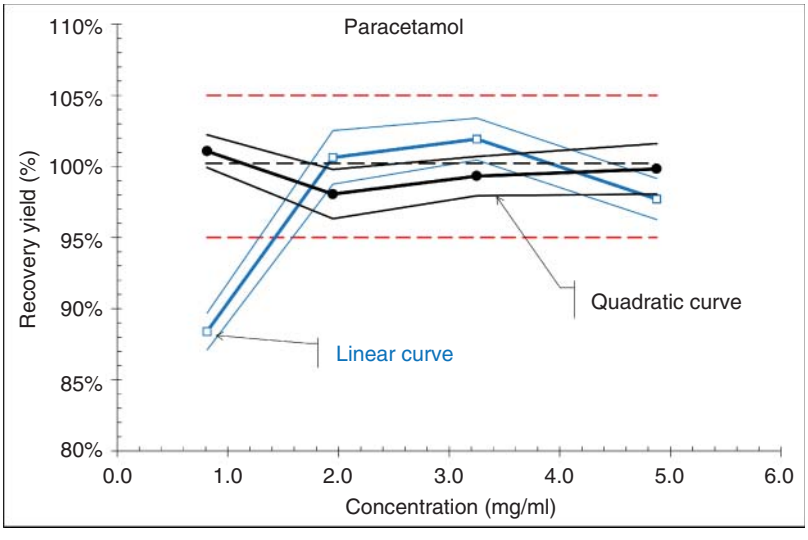



**Figure 8.9** PARACETAMOL – method accuracy profiles using two calibration models on the same data.

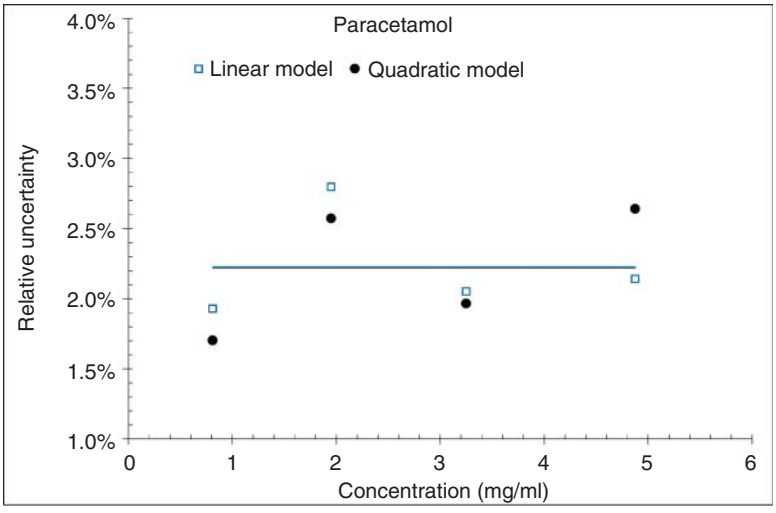



**Figure 8.10** PARACETAMOL – uncertainty functions for the two calibration models. White squares: linear, gray circles: quadratic model.

Considering MU, it is also possible to compute the corresponding values for each model and concentration with the statistical parameters already used to build the two profiles. These relative uncertainty values are illustrated in Figure 8.10 as a function of the paracetamol concentration. Since the validation range is narrow, relative uncertainty does not significantly vary. The most appropriate way to fit the relative uncertainty function is the straight line parallel to the concentration axis. Whatever the calibration model, linear or quadratic, both uncertainty functions overlap. This means that the relative MU is constant over the whole validated range and equal to 2.2% on average.

Finally, for the PARACETAMOL study, the calibration model has no influence on the MU. Several reasons may explain this fact: the quite high concentrations of the validation materials ranging from 1.0 to 5.0 mg/ml; by definition, MU only considers the dispersion of results and not the bias that is assumed corrected, while a flawed calibration model mainly impacts the bias.

This is not a general conclusion, and there are other situations where an inappropriate calibration model may alter the MU. The same comparison can be made with the THEOPHYLLINE dataset, where several uncertainty functions can be built with different calibration models.

Figure 8.11 compares the uncertainty functions derived from inverse-predicted results when the quadratic model is calculated with the OLS or WLS regression method. It seems logical to assume that the best model reduces the MU. This assumption can be verified here as the WLS model produces smaller MU, mainly at low concentrations. Several reasons can also be mentioned to explain this difference: the low concentrations of the validation materials ranging from 0.05 to 10.0 µg/l (i.e. $10^6$ times lower than paracetamol); the vicinity to LOQ; the daily variability of the calibration model coefficients, and so on.

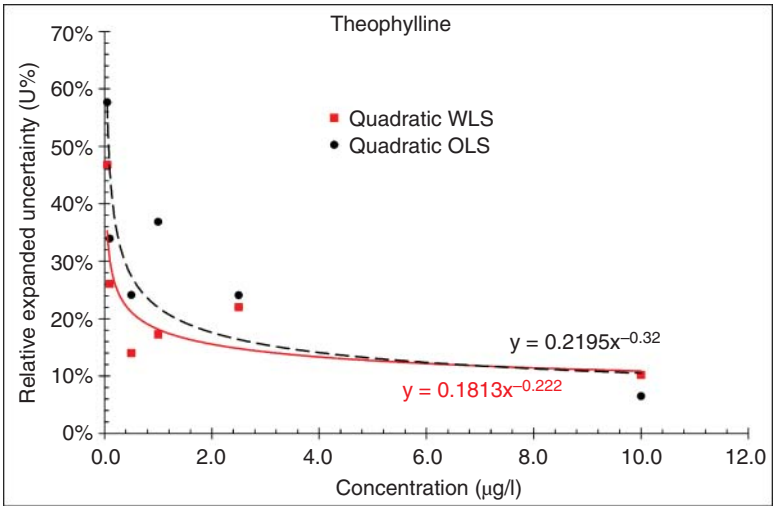



**Figure 8.11** THEOPHYLLINE – comparison of uncertainty functions obtained with two calibration models.

### 8.4.2 Uncertainty of Corrected Results

To fully apply the GUM general procedure, before providing an estimate of MU, it is stated in clause 3.2.4:

> "… it is assumed that the result of a measurement has been corrected for all systematic effects recognized as significant and that every effort has been made to identify them…"

This clause implies that, in most cases, a correction is possible by modifying the calibration function. However, if this is true for physical measurements, it is not often the case in the analytical sciences. The method accuracy profile (MAP) can be a tool for identifying the eventual bias of the method, and it is possible to graphically check if the trueness of the method is satisfactory.

When the bias is considered a significant source of inaccuracy, it is necessary to establish a correction factor applicable to the inverse-predicted concentrations. In this case, the uncertainty function should be modified accordingly. When considering the theophylline MAP illustrated by Figure 5.5 it is possible to identify a significant bias at low concentration. Table 5.4 reports the values of the relative bias as a function of the concentration.

Let us remember that the relative bias is equal to 100% – recovery yield. For instance, for a concentration of 0.1 µg/l the recovery yield is 88% and the relative bias 12%. The latter diminishes very quickly and becomes negligible in regard to the MU in the rest of the validation range.

THEOPHYLLINE: Relative bias as a function of concentration

| Levels (µg/l) | 0.05 | 0.1 | 0.5 | 1 | 2.5 | 10 |
|---|---|---|---|---|---|---|
| Relative bias (%) | 17 | 12 | 4 | 0 | 1 | 4 |

For this example, because it is estimated that the LOQ is about 0.19 µg/l, it seems unnecessary to correct the data for the relative bias. But, once the need to apply a correction factor has been identified and validated, it becomes a part of the operating procedure and must be applied systematically to all new measurements. This statement must be qualified by two remarks:

- The correction factor is specific to a given matrix.
- The correction factor may vary with the concentration level.

The correction factor is a new source of uncertainty that must be included in the estimate of the MU for the corrected measurement value. It should be remembered that correcting biased measurements is a perfectly acceptable practice within ISO 17025 accreditation. It must be scientifically justified and reported to the end-user.

Reviews of different methods applicable to the estimation of the MU of a corrected measurement are available in [15, 16]. The solution presented below is developed in agreement with the GUM recommendation described in Appendix F.2.4.5 of GUM (2018 version). It is stated that "when corrections are not applied from a calibration curve…" an approach described as "relatively simple for this problem, consistent with GUM principles" can be applied. The GUM procedure consists of calculating an average correction factor for a given measurand. Let us consider the data collected for building a MAP.

The measurand is the assigned target concentration $X_k$ of validation material. Following the proposed notation, $Z_{ijk}$ is the inverse-predicted concentration of replicate $j$ ($1 \le j \le J$) in series $i$ ($1 \le i \le I$) for the validation material $k$ ($1 \le k \le K$). The absolute individual bias $\delta_{ijk}$ is defined by Equation (8.9). The average correction factor for a given value of $X_k$ is defined by the arithmetic mean of all biases $\overline{CF_k}$. The corrected predicted concentration $Z^*_{ijk}$ is defined by Equation (8.11). It is essential to utilize an additive correction factor, whereas a multiplicative one would increase or decrease the measurement value and artificially modify the MU. The standard variance of the average correction factor is defined by Equation (8.12).

The bias of an inverse-predicted concentration

$$\delta_{ijk} = Z_{ijk} - X_k \tag{8.9}$$

Average correction factor for level $k$

$$\overline{CF_k} = \frac{\sum_i \sum_j \delta_{ijk}}{IJ} \tag{8.10}$$

Corrected measurement value

$$Z^*_{ijk} = Z_{ijk} - \overline{CF_k} \tag{8.11}$$

The standard variance of the average correction factor

$$u^2(\overline{CF_k}) = \frac{\sum_i \sum_j (\delta_{ij} - \overline{CF_k})^2}{(IJ - 1)} \tag{8.12}$$

Applying the law of propagation of uncertainty for a sum of input quantities, the combined standard uncertainty of a corrected measurement is:

The standard uncertainty of a corrected measurement

$$u\left(Z^*_{ijk}\right) = \sqrt{u^2(Z_{ijk}) + u^2(\overline{CF_k})} \tag{8.13}$$

The relative uncertainty of a corrected measurement

$$UR\%\left(Z^*_{ijk}\right) = \frac{u\left(Z^*_{ijk}\right)}{Z^*_{ijk}} \tag{8.14}$$

As the standard uncertainty $u(Z_{ijk})$ is computed with the uncertainty function derived from the accuracy profile while the standard uncertainty of $\overline{CF_k}$ is given by formula (8.12), this procedure is easy to implement. New values of standard and relative uncertainty are obtained to readjust the uncertainty function. The latter will be applicable to any new sample analyzed and corrected by the laboratory.

But an additional question remains when using this method routinely. How to determine the correction factor to be applied to any new sample result? When the GUM correction approach is applied to each distinct level of the accuracy profile, it appears that the obtained correction factors and associated uncertainties vary with the concentration. If a new unknown sample is not close to one of the $X_k$ value, an adjusted correction factor must be used.

The relationship between the concentration and the correction factor *CF* can be established and used to interpolate for any concentration. This must be preferably linear, since one of the validation requirements for analytical sciences, is that the trueness is proportional to the concentration, as explained in Section 4.1. Equation (8.15) is the mathematical interpretation of this requirement, where $p$ is the proportionality coefficient. This is the equation of a straight-line forced through the origin because when the concentration is 0, no correction is required:

Proportionality relationship between the correction factors and the concentration

$$CF = p \times X \tag{8.15}$$

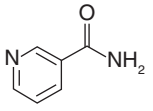
The coefficient $p$ is the slope of the curve and can be positive or negative. In Equation (8.15), the predicted correction factor is expressed in the same units as $X$ and can be applied directly to correct a value as in Equation (8.17). The next step is redrawing the initial MAP and obtaining a corrected profile. If it is satisfactory and the method validated, it can be considered as a verification of the relative correction factor as in equation (8.16). Another way consists in expressing the global correction factor as a percentage relative to the concentration. Then it applies to any recovery yield and makes it possible to compute the corrected accuracy profile.

Relative correction factor

$$\frac{CF}{X} \times 100 = p\% \tag{8.16}$$

Corrected measurement value

$$Z^* = Z \times (1 - p) \tag{8.17}$$



Nicotinic acid

To illustrate this method, the NICOTINIC dataset described below is used. It is recognized in the food industry to supplement foods with diverse nutrients. The supplementation of cow milk with vitamin B3 (nicotinamide) or nicotinic acid is widely applied. Nicotinic acid is more stable and easily converted to nicotinamide in vivo. Nicotinamide is the biological form; it is a derivative of nicotinic acid and a water-soluble vitamin.

Although both have the same vitaminic activities, nicotinic acid is the preferred form for industrial food supplementation. The NICOTINIC method consisted of the determination of nicotinamide and nicotinic acid in cow's milk. The corresponding dataset is described in Table 8.5. Only the determination of nicotinic acid is addressed because the determination of nicotinamide was not questionable. Original inverse-predicted concentrations for the three validation materials are listed in Table 8.6. In the same table individual and average biases are reported.

As described in Section 5.2.2, the MAP from this data can be established when using the Resource H Excel worksheet, only considering β-ETIs. Results are illustrated in Figure 8.12 by the accuracy profile labeled "Before correction." There is an obvious and strong systematic relative bias close to –50% that requires being corrected. The main parameters of the MAP that confirm this graphical interpretation are listed below:

| | | | |
|---|---|---|---|
| Concentration (mg/l) | 0.2 | 2 | 4 |
| Recovery yield (%) | 53 | 54 | 52 |
| Relative bias (%) | −47 | −46 | −48 |

The starting point of the correction procedure proposed by the GUM consists of calculating for each concentration level the average correction factor and its

**Table 8.5** NICOTINIC – description of the dataset.

| **Title** | **NICOTINIC** |
|---|---|
| Reference | Unpublished personal data |
| Measurand | Concentration of nicotinic acid in milk, expressed in mg/l. |
| Method | HPLC coupled to a fluorescence detector |
| Validation interval | [0.2, 4.0] mg/l |
| Acceptance interval | ± 10% |
| Validation materials | 3 validation materials containing 0.2, 2.0 and 4.0 µg/l respectively, prepared by SAM from a batch of homogenized milk. |
| Validation plan | Series ($I = 3$), replicates/series ($J = 3$), levels ($K = 3$) |
| Calibration plan | Series ($I' = 2$), replicates/series ($J' = 2$), levels ($K' = 2$) |
| Number of measures | 12 measurements on calibration solutions<br>27 measurements on spiked materials. |

**Table 8.6** NICOTINIC – original inverse-predicted concentrations and bias in mg/l.

| Level $X_k$ | Series | Inverse predicted concentrations | | | Individual bias | | | Correction factor | |
|---|---|---|---|---|---|---|---|---|---|
| | | $Z_{i1k}$ | $Z_{i2k}$ | $Z_{i3k}$ | $\delta_{i1k}$ | $\delta_{i2k}$ | $\delta_{i3k}$ | $\overline{CF_k}$ | $u^2(\overline{CF_k}) \times 10^{-3}$ |
| 0.2 | 1 | 0.096 | 0.113 | 0.109 | −0.104 | −0.087 | −0.091 | −0.093 | 0.598 |
| | 2 | 0.100 | 0.111 | 0.121 | −0.100 | −0.089 | −0.079 | | |
| | 3 | 0.100 | 0.105 | 0.105 | −0.100 | −0.095 | −0.095 | | |
| 2 | 1 | 1.167 | 1.090 | 1.043 | −0.833 | −0.910 | −0.957 | −0.922 | 5.084 |
| | 2 | 1.008 | 1.073 | 0.998 | −0.992 | −0.927 | −1.002 | | |
| | 3 | 1.141 | 1.004 | 1.181 | −0.859 | −0.996 | −0.819 | | |
| 4 | 1 | 2.196 | 1.933 | 2.087 | −1.804 | −2.067 | −1.913 | −1.934 | 11.839 |
| | 2 | 2.027 | 1.893 | 2.008 | −1.973 | −2.107 | −1.992 | | |
| | 3 | 2.171 | 2.178 | 2.102 | −1.829 | −1.822 | −1.898 | | |

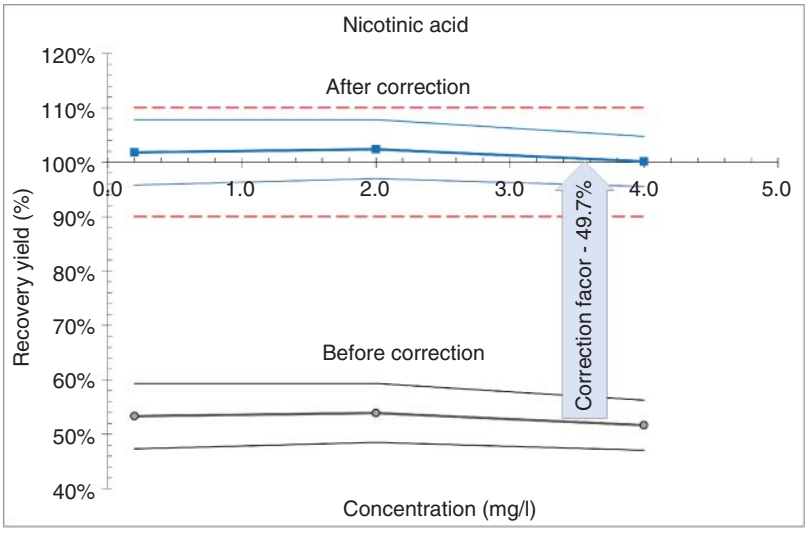



**Figure 8.12** NICOTINIC – accuracy profiles before and after correction.

standard variance following Equations (8.10) and (8.12). The values obtained are reported in the last columns of Table 8.6. The most convenient way to compute the proportionality coefficient $p$ is to apply the `LINEST` built-in function forcing the line to go through zero, i.e. setting the `Const` argument to `FALSE`. The value obtained is $p = -0.479$ expressed in mg/l; that gives $p\,\% = -47.9\%$ as a percentage.

While the initial accuracy profile is expressed as recovery yields, the corrected accuracy profile is obtained by subtracting this global correction factor from each value of the β-ETI bounds, themselves expressed in %. In Figure 8.12, the new

**Table 8.7** NICOTINIC – influence of the correction factor on the uncertainty and accuracy profile.

| | Parameters | Symbol | Concentration (mg/l) | | |
|---|---|---|---|---|---|
| | | | 0.2 | 2 | 4 |
| | | Accuracy profile | | | |
| Before | Recovery yield | | 53% | 54% | 52% |
| Before | Lower limit of β-ETI | | 47% | 49% | 47% |
| Before | Upper limit of β-ETI | | 59% | 59% | 56% |
| | Correction factor | $p\%$ | | −47.9% | |
| After | Recovery yield | | 101% | 102% | 100% |
| After | Lower limit of β-ETI | | 95% | 96% | 95% |
| After | Upper limit of β-ETI | | 107% | 107% | 104% |
| | | Measurement uncertainty | | | |
| Before | Standard uncertainty | $u(Z)$ | 0.0086 | 0.0767 | 0.1249 |
| Before | Relative uncertainty | $UR\%(Z)$ | 8.56% | 7.67% | 6.25% |
| Before | Standard variance | $u^2(Z)$ | $7.32{\cdot}10^{-5}$ | $5.88{\cdot}10^{-3}$ | $1.56{\cdot}10^{-2}$ |
| Before | Variance of the *CF* | $u^2(\overline{CF}_k)$ | $5.98{\cdot}10^{-5}$ | $5.08{\cdot}10^{-3}$ | $1.19{\cdot}10^{-2}$ |
| After | Standard uncertainty | $u(Z^*)$ | 0.01153 | 0.10472 | 0.16570 |
| After | Relative uncertainty | $UR\%(Z^*)$ | 11.53% | 10.47% | 8.29% |

accuracy profile is labeled "After correction." Table 8.7 table summarizes these calculations. Figure 8.12 perfectly illustrates how the global relative correction factor $p\%$ works, as the new corrected accuracy profile is now entirely within the acceptance interval limits.

Figure 8.13a shows the relationship between the correction factor and the concentration described by Equation (8.15). The linearity is established between the trueness and the concentration. However, the dispersion of individual biases around the average, symbolized by a small dashed line, increases when the concentration increases, and the WLS regression technique would improve the estimation of the slope of the straight-line. Figure 8.13a shows the relative uncertainty functions before and after correction.

These are power functions, as described in Section 7.5.2.

$$UR\%(Z) = 0.0753 \times Z^{-0.091}$$

$$UR\%(Z^*) = 0.1015 \times Z^{-0.093}$$

The function of the corrected measurements is shifted upwards, indicating that the uncertainty is positively impacted by the correction factor. This mainly affects the constant of the function, which increases from 0.0753 to 0.1015, i.e. a difference of about 0.025 that can be interpreted as 2.5% of increase in the relative uncertainty.

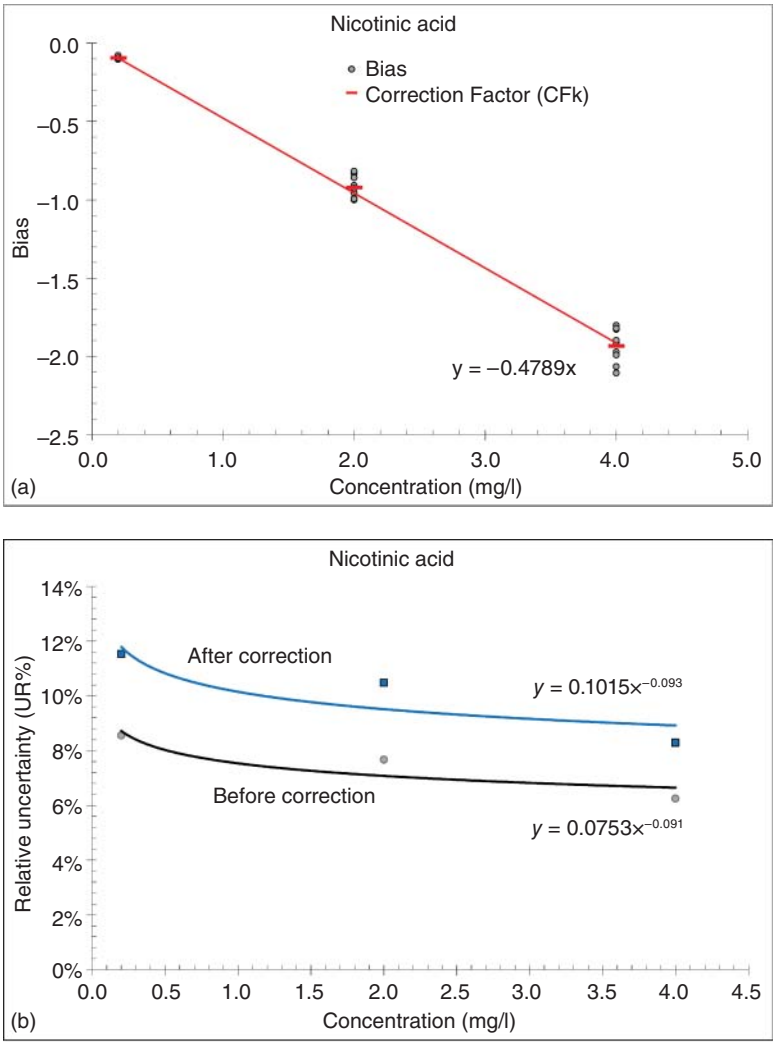



**Figure 8.13** (a) NICOTINIC – relationship between the average correction factor and the concentration. (b) NICOTINIC –Relative uncertainty function before and after correction.

On the other hand, the power coefficients of both functions can be considered as constant. The reporting of the correction factor uncertainty only generates a constant shift. In conclusion, the correction of the measurement values increases the MU. In this example, the global correction factor is constant and close to 50%.

For example, let us take a milk sample for which the observed raw concentration of nicotinic acid $Z$=0.80 mg/l. After applying the correction factor $p = -0.479$, the corrected concentration becomes $Z^{*}$=0.80 × (1–(–0.479)) = 1.18 mg/l, and the relative uncertainty is estimated at 9.8% with a coverage interval of [1.07, 1.31].

### 8.4.3 Increase the Number of Replicates

In some cases, the obtained MU is higher than expected and makes it difficult to decide with a sufficient level of confidence. A classic example is the compliance of a sample containing traces when the specification limit is close to the limit of quantification. A possible way to overcome this shortcoming is to work with replicates and make the decision based on an average result. This approach impacts the MU and implies some corrections for this way of expressing the result.

First of all, the term *replicate* needs to be defined. At what stage of the analytical operating procedure does the replication begin? When are different test portions sampled, when the extract (or the reconstituted sample) is injected into the measuring instrument, or at another moment? The replicates done at different starting points do not have the same uncertainty, since diverse sources of uncertainty may be triggered. The diagram presented in Figure 8.14 is an attempt to illustrate the multiplicity of replicate definitions and formalize this concept of replicate.

It is now accepted to divide an analytical process into three complementary parts: pre-analytical, analytical, and post-analytical. Replication itself is limited to the analytical part. The US-Pharmacopeia (USP) document [17] proposes also

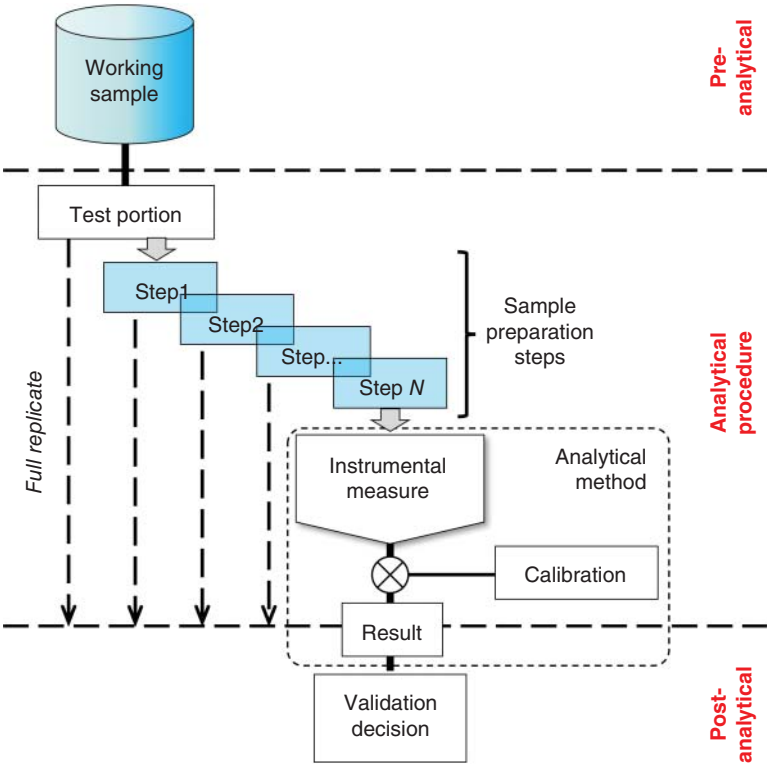



**Figure 8.14** Four possible replicate definitions according to the sample preparation starting step among the sample preparation steps.

to make a division, within the analytical part, between the analytical procedure, which includes the sample preparation steps, and the analytical method, which corresponds to the actual instrumental measurement, the calibration, and the result calculation.

This partition is helpful to better define replication strategies intended to reduce MU in drug monitoring [18]. In Figure 8.14 the vertical dashed arrows illustrate four distinct types of replicates that can be performed starting at various stages of a sample preparation. The replication procedure, labeled "*Full replicate*" on the figure, starts from a new test portion and consists of applying the complete analytical operating procedure. This is the most interesting because it represents the highest level of independence that can be obtained between two replicates done on the same working sample.

Unfortunatly, the word "sample" is often indiscriminately and carelessly used to refer either to the entity that is received (or collected) by the laboratory for analysis or to the resulting preparation, such as an extract or reconstituted material after sample preparation, which is introduced into the measuring instrument. In the previous Section, 8.3, about sampling uncertainty, some definitions were proposed to avoid this confusion. It is essential that all replicates be obtained on the same sample, i.e. the same entity received by the laboratory.

Let us note $Z_n$ with $1 \leq n \leq N$ the replicates used to calculate an average result. Two situations must be clearly distinguished while in each case, the result MU is differently estimated when dealing with the average of the replicates:

- Measures under repeatability condition. Replicates are performed on the same day by the same operator in the same series. It can be assumed they have the same standard uncertainty, denoted $u(Z_n)$.
- Measures under intermediate precision conditions. A common way to make replicates when a sample is declared nonconforming to check the consistency of the decision is to repeat one measurement (or several measurements) later during another sequence or series.

### 8.4.4 Replication under Repeatability Condition

A reliable replicate should be performed from a new test portion, as described in Figure 8.14 to include as many sources of uncertainty as possible under repeatability conditions. It is also the simplest situation to compute the result MU. If the uncertainty function of the method has been established as described in Section 7.5, it is possible to estimate $u(Z_n)$ for any $Z_n$. According to the GUM, two measurement models should be considered for an average of replicates. They are denoted here $\overline{Z_A}$ and $\overline{Z_B}$.

Model A

$$\overline{Z_A} = \frac{\sum_{n=1}^{N} Z_n}{N} \tag{8.18}$$

Model B

$$\overline{Z_B} = \frac{\sum_{n=1}^{N} Z_n}{N} + E_r \tag{8.19}$$

Model B includes the idea that measurements, although obtained under repeatability conditions, are not all equal, and a residual random error accounts for the differences. The random variable $E_r$ reflects these random variations between replicates. When these models are developed according to the law of uncertainty propagation, the following MU estimate is obtained for each model.

Standard variance of model A

$$u^2(\overline{Z_A}) = \frac{\sum_{n=1}^{N} u^2(Z_n)}{N} \tag{8.20}$$

Standard uncertainty of model A

$$u(\overline{Z_A}) = \sqrt{u^2(\overline{Z_A})} \tag{8.21}$$

Standard variance of $E_r$

$$u^2(E_r) = \frac{\sum_{n=1}^{N} (Z_n - \overline{Z_A})^2}{\frac{N-1}{N}} \tag{8.22}$$

Standard uncertainty of model B

$$u(\overline{Z_B}) = \sqrt{u^2(\overline{Z_A}) + u^2(E_r)} \tag{8.23}$$

A classic statistical property is used here, namely: if a random variable is the sum of $N$ random variables of the same standard deviation $\sigma$, its standard deviation is $\frac{\sigma}{\sqrt{N}}$. This property allows us to set the following table, which shows the possible reduction gain on the MU when the number of replicates increases from one single measure. For example, to reduce MU by one-half, it is necessary to do at least four replicates.

| $N$ | $1/\sqrt{N}$ | Gain (%) |
|---|---|---|
| 2 | 0.71 | 29 |
| 3 | 0.58 | 42 |
| 4 | 0.50 | 50 |
| 5 | 0.45 | 55 |
| 8 | 0.35 | 65 |
| 10 | 0.32 | 68 |

The results of the THEOPHYLLINE study will serve as an example. The values obtained for eight replicates made on the same sample are reported in Table 8.8. The column "Mean" corresponds to the provisional calculation of the average; for example, 0.1515 corresponds to 2 replicates. The standard uncertainty function of the method is provided in Section 7.5.2 and illustrated in Figure 7.8a. It is a power

**Table 8.8** THEOPHYLLINE – influence of the number of replicates.

| | | | Model A | | Model B | |
|---|---|---|---|---|---|---|
| **Measure number** | **Value** | **Mean** | $u^2(\overline{Z_A})$ | $u(\overline{Z_A}) \times 10^3$ | $u^2(e_r)$ | $u(\overline{Z_B}) \times 10^3$ |
| 1 | 0.154 | | | 21.150 | | 21.150 |
| 2 | 0.149 | 0.1515 | $2.17 \cdot 10^{-4}$ | 14.728 | $6.25 \cdot 10^{-6}$ | 14.939 |
| 3 | 0.151 | 0.1513 | $1.45 \cdot 10^{-4}$ | 12.025 | $2.11 \cdot 10^{-6}$ | 12.113 |
| 4 | 0.155 | 0.1523 | $1.08 \cdot 10^{-4}$ | 10.414 | $1.90 \cdot 10^{-6}$ | 10.505 |
| 5 | 0.150 | 0.1518 | $8.68 \cdot 10^{-5}$ | 9.315 | $1.34 \cdot 10^{-6}$ | 9.387 |
| 6 | 0.148 | 0.1512 | $7.23 \cdot 10^{-5}$ | 8.503 | $1.29 \cdot 10^{-6}$ | 8.579 |
| 7 | 0.152 | 0.1513 | $6.20 \cdot 10^{-5}$ | 7.873 | $9.39 \cdot 10^{-7}$ | 7.932 |
| 8 | 0.151 | 0.1513 | $5.42 \cdot 10^{-5}$ | 7.364 | $7.05 \cdot 10^{-7}$ | 7.412 |

function with the following coefficients: constant $a = 0.0907$ and power $b = 0.778$. Thus, the standard uncertainty of the first measure 0.154 µg/l is $u(Z)$=21.150. $10^{-3}$ µg/l.

All MU values are multiplied by $10^3$ to be more readable. As expected, Table 8.8 shows that the MU of the average result regularly decreases from 21. $10^{-3}$ to 7.$10^{-3}$, when the number of replicates increases from 1 to 8, i.e. a final 67% reduction. Obviously, the cost of this procedure can be high, and a cost/benefit study may help to decide the adequate replicate number. When comparing the estimates obtained with model A or B, the correction introduced in model B only slightly modifies the MU estimate.

Further refinement can be introduced in the measurement model of a result corresponding to an average of replicates because replicates are made under repeatability conditions and consequently correlated. The law of propagation of uncertainty can consider this kind of correlation as illustrated by Equation (6.18) before it is simplified in Equation (6.21). The difficulty is to evaluate practically this correlation.

### 8.4.5 Replication under Intermediate Precision Condition

As previously explained, when a measurement value is not as satisfactory as expected (e.g. when a whole production would not be marketable because of one nonconforming result), it is quite common to do a replicate to infirm, confirm, strengthen, or weaken the conclusion. Sometimes, this operation is repeated several times to achieve a certain degree of confidence for the decision. If the new analysis does not belong to the same series, the between-series effect must be added, which implies that an estimate is available.

To explain the calculation, the usual notation is utilized. Let us assume that $n$ replicates are done; the uncertainty function is available, and an estimate $u(Z_n)$ is available for each. This results in the following formulas.

Mean (result)

$$\overline{Z} = \frac{\sum_{n=1}^{N} Z_n}{I} \tag{8.24}$$

Within-replicate variance

$$s_r^2 = \frac{\sum_{n=1}^{N} u^2(Z_n)}{N} \tag{8.25}$$

Between-replicates variance

$$s_B^2 = ? \tag{8.26}$$

Standard variance of the mean

$$u^2(\overline{Z}) = \frac{s_B^2 + s_r^2}{N} \tag{8.27}$$

This is an uncomfortable situation when no estimate of the between-replicates variance is available, but it is common. Therefore, it is a risky practice to do replicates under intermediate precision and compute an average to make a final decision. Decision rules come with a sampling plan. If only one measurement is required to decide whether a lot is conforming, this rule must be respected. It is always possible to perform confirmatory analyses, but it is not correct to use the average for the decision unless special regulations exist.

## References


1 Travison, T.G., Vesper, H.W., Orwoll, E. et al. (2017). Harmonized reference ranges for circulating testosterone levels in men of four Cohort Studies in the United States and Europe. *Journal Clinical Endocrinology Metabolism* 102 (4): 1161–1173.

2 Salamin, O., Ponzetto, F., Cauderay, M. et al. (2020). Development and validation of an UHPLC–MS/MS method for extended serum steroid profiling in female populations. *Bioanalysis* https://doi.org/10.4155/bio-2020-0046.

3 2000/657/EC: Commission Decision of 16 October 2000 adopting Community import decisions for certain chemicals pursuant to Council Regulation (EEC) No 2455/92 concerning the export and import of certain dangerous chemicals.

4 Feinberg, M., Bertail, P., Tressou, J., and Verger, P. (2006). *Analyse des risques alimentaires*. Cachan: Lavoisier (in French).

5 FAO/Codex Alimentarius. Codex Pesticides Residues in Food Online Database. https://www.fao.org/fao-who-codexalimentarius/codex-texts/dbs/pestres/en/ (accessed 5 September 2023).

6 Regulation (EC) No 470/2009 of the European Parliament and of the Council of 6 May 2009 laying down Community procedures for the establishment of residue limits of pharmacologically active substances in foodstuffs of animal origin.

7 BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML (2012). Evaluation of measurement data: the role of measurement uncertainty in conformity assessment. JCGM 106. Sèvres, France. https://www.bipm.org (accessed 3 September 2023).

8 ISO/IEC Guide 98-4:2012. *Uncertainty of measurement — Part 4: Role of measurement uncertainty in conformity assessment*. ISO, Genève.

9 Williams, A. and Magnusson, B. (ed.) (2021). ISBN 978-0-948926-38-9. www.eurachem.org (accessed 3 September 2023). *Eurachem/CITAC Guide: Use of Uncertainty Information in Compliance Assessment*, 2e. EURACHEM-CITAC.

10 Commission Implementing Regulation (EU). 2021/808 of 22 March 2021 on the performance of analytical methods for residues of pharmacologically active substances used in food-producing animals and on the interpretation of results as well as on the methods to be used for sampling and repealing Decisions 2002/657/EC and 98/179/EC.

11 Heydorn, K. and Esbensen, K. (2004). Sampling and metrology. *Accreditation Quality Assurance* 9: 391–396.

12 Gy, P.M. (1982). *Sampling of Particulate Materials, Theory and Practice*. New York: Elsevier Scientific Publishing Company.

13 Grøn, C., Bjerre Hansen, J., Magnusson, B. et al. (2007). Uncertainty from Sampling – A NORDTEST Handbook for Sampling Planners on Sampling Quality Assurance and Uncertainty Estimation. *Technical Report 604*, NORDTEST.

14 Ibrahim, A.M., Hendawy, H.A.M., Hassan, W.S. et al. (2019). Data on validation using accuracy profile of HPLC-UV method. *Data in brief* 24: 103877.

15 Maroto, A., Boqué, R., Riu, J., and Rius, F.X. (2001). Measurement uncertainty in analytical methods in which trueness is assessed from recovery assays. *Analytica Chimica Acta* 440: 171–184.

16 Vanatta, L.E. and Coleman, D.E. (2007). Calibration, uncertainty, and recovery in the chromatographic sciences. *Journal of Chromatography A* 1158: 47–60.

17 Schofield, T., van den Heuvel, E., Weitzel, J. et al. (2020). Distinguishing the Analytical Method from the Analytical Procedure to Support USP Analytical Life Cycle Paradigm. *USPF Online*.

18 Borman, P., Schofield, T., and D. (2021). Lansky: reducing uncertainty of an analytical method through efficient use of replication. *Pharmaceutical Technology* 48–56.

# 9

# MU and Quantification Limits

The ability of a method to quantify working samples at low analyte concentration is an important topic of discussion because once a limit is established, it should facilitate method selection. Very often, it is also used as a commercial argument. Most commonly, the quantification capacity threshold is defined through two parameters, the limit of detection (LOD) and the limit of quantification (LOQ), even though many other parameters are also proposed in the analytical literature.

LOD is defined by metrologists as "the measured value, obtained by a given measurement procedure, for which the probability of falsely claiming the absence of a component in a material is, given a probability of falsely claiming its presence" (International Vocabulary of Metrology [VIM]). It applies to any measuring method and is not limited to analytical sciences.

This definition is compatible with statistical significance testing that involves two risks of error, usually named $\alpha$ and $\beta$ (in this context, $\beta$ must not be confused with the $\beta$ probability of the β-expectation tolerance interval β-ETI). These error risks are classically fixed in the statistical literature at 1% or 5%, respectively.

It should be recalled that *detecting* an analyte leads to a qualitative result expressed by presence/absence, conforming/nonconforming, or yes/no. LOD theoretically represents the starting point where the instrumental signal is perceptible but not translated into a quantity of matter.

On the other hand, LOQ is specific to analytical sciences and quantitative methods. Unfortunately, there is no single standardized definition, either in VIM or in International Union of Pure and Applied Chemistry (IUPAC). It would rather correspond to an observed parameter derived from various extrapolations or interpolations. For quantitative methods, the real issue is the LOQ. It is therefore, this performance limit that is discussed here.

For qualitative methods, one of the parameters introduced to evaluate a method's performance limit is the equivalency of the level of detection (LD) accompanied by a probability (i.e. for example, LD50). Unfortunately, this abbreviation LD can lead to confusion with the LOD.

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

## 9.1 Definitions and Assessment of LOQ

For the LOQ, a definition and a calculation method were already proposed in Section 5.2.2, many other definitions exist and, consequently, many estimation methods. An attempt at a rough general definition can be proposed, and several variations are available in the guides, as explained in a recent review [1]. The most classical definition is the following, and in square brackets, suggested details are added:

> "The lowest concentration of an analyte that can be quantified [in a specific matrix] with acceptable precision and accuracy [and under specified operating conditions]."

But also:

> "The concentration for which the risks of error $\alpha$ and $\beta$ are acceptable."

This creates confusion with the LOD when the risk levels are not clearly specified. Ultimately:

> "The value for which the bias and relative standard deviation [or coefficient of variation] of repeatability do not exceed a given acceptance threshold."

From a methodological point of view, various calculation methods have been proposed. A succinct description may help to understand the practical problems raised by each procedure. Some decades ago, a well-documented historical review on the LOQ parameter was published but, unfortunately, not recently updated [2].

### 9.1.1 Multiple Blank Standard Deviations

In the 1970s, blank standard deviation was the method recommended by the Environmental Protection Agency (EPA). The simplest solution is to calculate the LOD and LOQ as $k$-fold standard deviation of the intercept $a_0$ of the calibration curve (assumed being a straight line) denoted $s_{a_0}$ and given by Eq. (2.18). This parameter is assumed to be sufficiently comparable to the standard deviation $s_{bl}$ of a blank sample.

Whatever the $k$-fold value, the result is expressed in the same units as the instrumental signal. For instance, the peak area must be converted into a concentration by dividing by the slope $a_1$ (which corresponds to the sensitivity).

A classic methodological error is the use of a calibration curve that would neglect the matrix effects, the influence of the selected calibration model, or the possible heterogeneity of variances of the measurements at the different concentrations (or heteroscedasticity). Inverse calibration is a well-known operation to convert an instrumental response into a concentration.

Because the coefficient $a_1$ is an estimation of the slope, it may vary from one series to another, and computed LOD or LOQ may also vary. Moreover, if the intercept $a_0$ is significantly different from zero (or even negative), another source of variability

is present, which is not always easy to consider. For the LOD, a multiplicative coefficient $k_D$ is applied to the standard deviation $s_{a_0}$. It varies according to the validation guide, for example $k_D = 3.0$, 3.3, 6.0, etc. These values approach Normal distribution law quantiles used in statistical significance tests but do not strictly conform to the theory.

$$\text{LOD} = k_D \times \frac{s_{a_0}}{a_1}$$

For LOQ, a similar principle remains, and by convention, the EPA proposed to take a multiplicative coefficient $k_Q = 10$, without any statistical justification, just according to a general principle that implies that the LOQ must be "sufficiently distant" from the LOD. Finally, this approach often leads to unrealistic values but is nevertheless accepted by control authorities.

$$\text{LOQ} = k_Q \times \frac{s_{a_0}}{a_1}$$

### 9.1.2 Visual Examination

Accepted in the pharmaceutical industry [3] and in biomedical analysis, this method relies on analyst's expertise who knows the discrimination capability of the method he applies to expect a correct quantification. The justification of the threshold is provided by signal records, such as a chromatogram. In this case, the LOQ is directly expressed in the unit of the analyte, i.e. in absolute or relative concentration. The drawback is the subjective aspect of the procedure. If an obtained value is used for commercial promotion, it may be uneasily arguable.

### 9.1.3 Signal-to-Noise Ratio

This approach is based on the physical theory of instrumental signals. It mainly applies to spectroscopic methods, such as atomic emission, gamma-ray spectrometry, nuclear magnetic resonance, ultraviolet (UV)–Visible. The procedure often advocated is to compare the signals obtained with low-concentration samples to those from blanks and determine the minimum concentration for which the analyte can be detected or quantified.

A signal-to-noise ratio $S/N = 3:1$ is generally considered acceptable for a LOD and 10:1 for a LOQ. The same values as previously assigned to $k_D$ and $k_Q$. It should be noted that in the context of separative methods, quantification is often carried out based on integrated peak area. In this case, a LOQ obtained from the signal-to-noise ratio does not make sense because the word "ratio" does not refer to the area but to the signal height.

Therefore, this approach has serious drawbacks with modern instrumentation, where the signal is digitized. The baseline signal is usually filtered and smoothed, and it is difficult, if not impossible, to measure a true instrumental background. Hence, the evolution of recent instruments, such as high-resolution mass spectrometers, has resulted in extremely low background noise, and this approach has become pointless. Some discussion about the inconsistency of the signal-to-noise ratio with this kind of method is available in [4].

### 9.1.4 Empirical Experimental Approach

This consists of repeated measurements on different test portions of a blank sample (i.e. analyte-free) or as close as possible to the supposed LOQ. This approach has the advantage of directly starting from measured data. Besides, it raises enormous practical problems, such as: how to obtain such a sample; how to deal with the qualitative but non-quantifiable data that appear around the LOQ, etc.? In conclusion of this rapid overview of the various assessments of the LOQ, it is obvious that there is no consensus.

All these procedures provide values that could be quite different. Several studies have demonstrated that LOQs obtained from identical data, but calculated according to different procedures, can be highly divergent, as explained in [2]. Moreover, it is common to obtain the value by extrapolation outside the validation domain. Indeed, in the vicinity of the LOQ, it seems logical to obtain nonquantifiable instrumental responses that cannot, by definition, be treated by classic statistical methods. Specific tools applicable to this *censored data*, mixing qualitative and quantified data, exist and are available.

This ambiguity around LOQ definitions and estimators may be surprising, considering the importance given by analysts to this parameter. It is undoubtedly imperative to establish a consensus. In any case, the recommendation to validate the LOQ rather than estimate it seems not only reasonable but mandatory. This confirmation can be done by means of surrogate samples and/or samples spiked by standard additions, then submitted to the whole analytical procedure.

All the above approaches are ultimately only estimates of the LOQ. Finally, it should be noted that in some international regulatory texts, a lower limit of quantification or LLOQ and, by symmetry, an upper limit of quantification or ULOQ are defined and required [5]. It is more a way to define the practical bounds of the validated domain rather than the whole method performance itself.

## 9.2 LOQ as an Expected Relative Uncertainty

The previously reported LOQ definitions and the proposed calculation procedures do not consider the measurement uncertainty (MU). Moreover, if the same experimental design is applied several times in a row, different LOQ values would be obtained. Based on this observation, another definition of LOQ that includes the MU can be proposed:

> "LOQ is the lowest concentration of an analyte that can be quantified in a given matrix with a defined MU that could be absolute or relative (e.g., 50%) under specified operating conditions."

The great interest of the uncertainty function is to predict the MU for a given measurement. When the inverse uncertainty function is taken, it can also be used to predict a concentration for a given MU value, as shown in the following formulas and explained in the Section 7.6.

**Table 9.1** Different LOQ values calculated by inverting the uncertainty function from the THEOPHYLLINE study (µg/l).

| | Method accuracy profile | Relative uncertainty values *UR%* | | | | |
|---|---|---|---|---|---|---|
| | | 80% | 60% | 50% | 30% | 20% |
| LOQ (µg/l) | 0.129 | 0.001 | 0.005 | 0.01 | 0.104 | 0.643 |

Power uncertainty function

$$UR\% = c \times X^{d}$$

Uncertainty function THEOPHYLLINE

$$UR\% = 0.8968 \times X^{-0.271}$$

Inverse uncertainty function

$$X = 10^{\left(\frac{\log(UR\%)-\log(c)}{d}\right)} \tag{9.1}$$

Inverse function for THEOPHYLLINE

$$X = 10^{\left(\frac{\log(UR\%)-0.0473}{-0.271}\right)}$$

Applying this method to the results of the THEOPHYLLINE study, it is easy to construct Table 9.1, which brings together the predicted $X$ concentrations for different values of $UR\%$. According to the new definition, each example is a possible LOQ value that can be compared to the value obtained from the accuracy profile as proposed in Section 5.2.2.

In this case, the analyst was questioned about the most *likely* LOQ value and answered for a relative uncertainty of 50% corresponding to a LOQ of 0.01 µg/l. This value is obtained thanks to an extrapolation, as the lowest experimental $X$-value is 0.05 µg/l. Although unacceptable for certain guidelines, it means that within the coverage interval [0.005, 0.015] µg/l would lie the 95% proportion of possible values of a LOQ-sample. It can also be noted that the value obtained from the accuracy profile is extremely far from this proposal since 0.129 µg/l corresponds to a relative MU of about 29% and a coverage interval of [0.092, 0.141] µg/l.

Hence, the accuracy profile (MAP)-derived approach better reflects the actual performance of the method because the LOQ is located inside the validated interval while it can be outside when derived from the uncertainty function. Another strategy is also possible, such as *a priori* defining the LOQ to be reached, computing associated relative uncertainty and coverage interval, and deciding whether it is appropriate or not.

While the MAP approach only provides an estimate of the LOQ, many regulatory bodies or official guidelines sometimes strictly require both parameters, LOD and LOQ. In this context, the question of estimating the LOD remains. Figure 9.1 is an attempt to illustrate a viable alternative solution.

With the classic approach (i.e. bottom-up), the computation starts with a blank standard deviation σ from which LOD and LOQ are deduced. As already stated, this

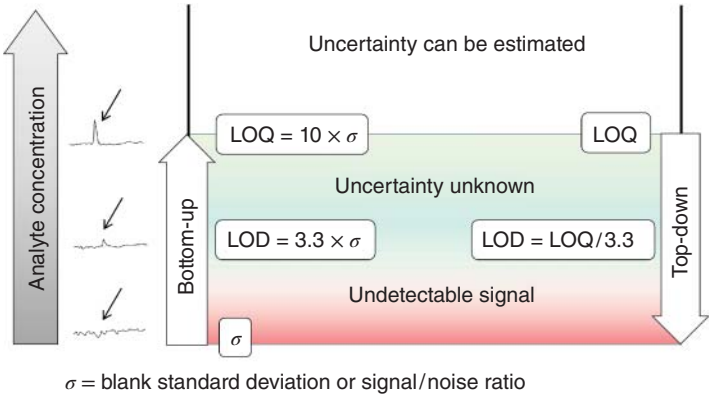



**Figure 9.1** Plausible classic and alternative approaches to estimate LOD and LOQ.

parameter has some physical meaning only when working with spectrophotometric techniques. Thus, a viable alternative (i.e. top-down) is first to consider the LOQ (derived from the MAP) and then deduce the LOD. When dealing with other techniques where the signal is a peak area and/or the background noise is extremely low, such as many chromatographic methods or MS quantification, the blank standard deviation is meaningless and nonmeasurable. The estimation of LOD based on signal-to-noise ratio becomes pointless [4].

## 9.3 Decision Limit and Detection Capability

### 9.3.1 Concepts and Definitions

LOQ is often provided by equipment manufacturers or analysts to assess the performance of analytical instrumentation or method. However, according to the previous paragraphs, it is obvious that the reported values can significantly vary, as a function of the calculation procedure. This raises a question of legitimacy and applicability. If the goal is to estimate to what extent a method can quantify the lowest possible concentration, the LOQ may be the appropriate parameter.

Paradoxically, it is probably not the most interesting objective from a practical point of view since it is a question of confirming that a method is capable of quantifying concentrations in accordance with the required specifications that have been established beforehand. This goal is to verify the adequacy of the method to the use that one wants to make of it, also called fitness-for-purpose. This property of a method was introduced in the ISO 11843:2000 standard under the name Capability of Detection. The latter plays a very practical role in selecting methods able to perform the official control of foods or the environment and has been introduced in several regulatory texts.

The evolution of European regulation on food hygiene can be used to illustrate how the concept of capability of detection works and has changed over recent decades. In 1996, the European Commission published a directive "on measures

to monitor certain substances and residues thereof in live animals and animal products" [6]. But it was not until 2002 that the practical implementation of the directive was published as a Decision [7].

For the latter, the traditional statistical approach based on null hypothesis testing inspired by the LOD definition was considered. More recently, the official procedure has moved to the use of the MU and is now comparable to the procedure described in Section 8.2 about sample conformity assessment. In 2002, two parameters related to the method detection ability were introduced, namely, $CC\alpha$ and $CC\beta$, derived from the concepts developed in the ISO 11843:2000 standard [8]. In 2002, the regulatory definitions were:

| | |
|---|---|
| *Decision limit* $CC\alpha$ | "The limit at which and beyond which it is permissible to conclude with a probability of error $\alpha$ that a sample is noncompliant." |
| *Detection capability* $CC\beta$ | "Smallest level of substance that can be detected, identified, and/or quantified in a sample with a probability of error $\beta$." |

In practice, the detection capability is presented as a threshold of non-compliance, with a risk of error fixed in advance. For its calculation, the official text distinguishes between:

- The "substances for which no permitted limit has been set," and the detection capability is comparable to the LOQ since it is the lowest concentration at which truly contaminated samples are detected.
- The "substances for which a permitted limit is set," the detection capability is the concentration at which the method can identify samples at the set limit.

The reference in the definition to the probability of error $\alpha$ or $\beta$ shows that it is the same logic as null hypothesis testing. The probability of error $\beta$ here-mentioned must not be confused with the proportion $\beta$ used to construct the β-expectation tolerance interval, abbreviated so far as β-ETI, and discussed in Chapter 5.

According to European legislation, two situations must be distinguished:

- The substance is prohibited and, therefore, must be absent.
- The substance is authorized but must not exceed a limit which can be a maximum residue limit (*MRL*), maximum level (*ML*), or "other tolerance applicable to substances" laid down in other Community legislation.

Figure 9.2 illustrates both officially defined parameters in the case of an authorized substance for which an *MRL* of 100 mg/kg. In this case, the required error probabilities are $\alpha = 5\%$ and $\beta = 5\%$, respectively [6].

From these probabilities, it is possible to define a coverage factor equal to 1.64 derived from the standardized normal distribution and applied to the standard deviation measured at the *MRL* and denoted $s_{MRL}$ and the standard deviations applicable to a sample $s_{sample}$ which is located "as close as possible to the *MRL*."

The shortcomings of this approach are obvious to an analyst and have already been pointed out above in relation to several LOQ definitions. They can be summarized as follows:

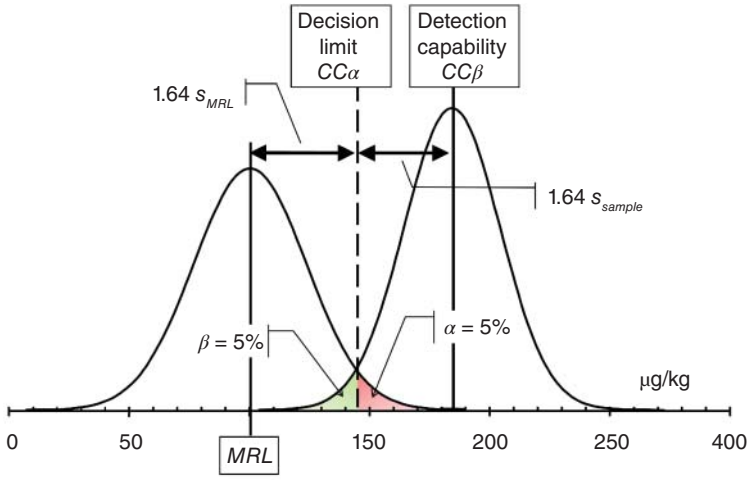



**Figure 9.2** Definitions of decision limit and detection capability according EU regulation [7] for a substance with a maximum residue limit (*MRL*) of 100 µg/kg.

- Experimental drawbacks. How can these standard deviations be estimated? How many replicates are needed? Is it necessary to be under repeatability or intermediate precision conditions? To answer these questions globally and because these data are unknown, it is classical to use, $s_{MRL} = s_{sample}$
- Response *versus* concentration. The starting point for the proposed definitions is the instrumental response. The conversion to a concentration involves using a calibration curve, as proposed by the EPA. However, the coefficients of the curve vary from one series to another, and this variability is not considered.

For all these reasons, in 2021 the European Commission published a new regulation on "the performance of methods of analysis for residues of pharmacologically active substances used in food-producing animals and the interpretation of results and the methods to be used for sampling and repealing Decisions 2002/657/EC and 98/179/EC" [9]. In this text, the parameter names and definitions have slightly changed. In particular, new calculation methods have been introduced even if, for regulatory compatibility, the approach defined in 2002 remains applicable.

| | |
|---|---|
| Decision limit for confirmation $CC\alpha$ | "The limit at which it is safe to conclude with a probability of error $\alpha$ that a sample is non-compliant, with the value $1-\alpha$ denotes the statistical certainty in percent that the allowable limit has been exceeded." |
| Detection capability for screening purposes $CC\beta$ | "The smallest analyte content that can be detected or quantified in a sample with a probability of error $\beta$." In the case of prohibited or unauthorized pharmacologically active substances, the $CC\beta$ is the lowest concentration at which a method can detect or quantify, with certainty $1-\beta$, samples containing residues of prohibited or unauthorized substances. For permitted substances, the $CC\beta$ is the concentration at which the method detects concentrations below the permitted limit with a statistical certainty of $1-\beta$. |

### 9.3.2 Initial Procedure (2002)

The first published estimator is based on the ISO 11843-2:2000 standard; it is known as the calibration curve method. The following formulas apply to any specification limit (*SL*) selected by the regulator. The parameter $k_\alpha = t_{1-\alpha,\nu}$ corresponds to the quantile of Student's $t$ law for the probability $\alpha$ and the number of degrees of freedom $\nu$. In an equivalent way $k_\beta = t_{1-\beta,\nu}$ is the quantile for the probability $\beta$.

Intermediate quantity

$$IQ = \frac{s_e}{a_1} \times \sqrt{\frac{1}{J} + \frac{1}{IJ} + \frac{(SL - \overline{X})^2}{SCE_X}} \tag{9.2}$$

Decision limit

$$CC\alpha = SL + k_\alpha \times IQ \tag{9.3}$$

Detection capacity (approximation)

$$CC\beta = CC_\alpha + k_\beta \times IQ \tag{9.4}$$

The parameter *SL* refers to any official compliance specification and takes different names depending on the regulatory text: *MRL*, maximum content (*MC*), reference value, and so on. Parameters $I$ and $J$ represent the number of calibrators and the number of replicates per calibrator, respectively. These notations differ from those used so far for the accuracy profile.

The so-called "intermediate quantity" *IQ* was introduced to facilitate the verification of a worksheet developed to calculate the parameters (see Chapter 10). In addition, it is possible to calculate the previously defined Critical response value denoted $Y_0$. It is the smallest response that allows for decision-making; it can be compared to the LOD before it is converted into concentration.

Critical response value

$$Y_0 = a_0 + k_\alpha \times s_e \sqrt{\frac{1}{J} + \frac{1}{IJ} + \frac{\overline{X}^2}{SCE_X}} \tag{9.5}$$

In these formulas, $a_0$ and $a_1$ represent the coefficients of the calibration curve (which is assumed to be a simple linear regression model), the intercept or blank, and slope or sensitivity, respectively and $s_e$ the residual standard deviation (see Section 2.2).

### 9.3.3 Modified Procedure (2021)

In the 2021 regulation update, a new and remarkably interesting calculation procedure was introduced, comparable to the principles of ISO Guide 98-4 on the declaration of conformity when considering the MU; this procedure is described in Section 8.2.

It consists in calculating a rejection zone by combining the regulatory specification limit *SL* and the MU of the method obtained by the laboratory at this concentration. The main difference with the ISO 98-4 guide is that the latter introduces the idea of

**Table 9.2** Varied procedures of calculating *CCα* and *CCβ*.

| *n°* | Principle of the procedure | Equations | |
|---|---|---|---|
| *Limit of decision for the purpose of confirmation (CCα)* | | | |
| Prohibited substances | | | |
| 1 | Calibration curve method according to ISO 11843-2 | Equations (9.2) and (9.3) with $SL = VR$ or $MC$ and $\alpha = 0.01$ | |
| 2 | 20 blanks and calculation of the signal-to-noise ratio *S/N* | $CC\alpha = 3.0 \times {}^{S}/_{N}$ | |
| 3 | Measurement uncertainty method | $CC\alpha = LCL + k_{\alpha=0.01}\, u_c(LCL)$ | (9.6) |
| Authorized substances | | | |
| 4 | Calibration curve method according to ISO 11843-2 | Equations (9.2) and (9.3) with $SL = MRL$ or $MC$ and $\alpha = 0.05$ | |
| 5[a] | Measurement uncertainty method | $CC\alpha = LMR + k_{\alpha=0.05}\, u_c(LMR)$ | (9.7) |
| 5[b] | | $CC\alpha = TM + k_{\alpha=0.05}\, u_c(TM)$ | (9.8) |
| *Detection capability for screening purposes (CCβ)* | | | |
| Prohibited substances | | | |
| 6 | Calibration curve method according to ISO 11843-2 | Equation (9.4) with $\beta = 0.05$ | |
| 7 | Measurement uncertainty method | $CC\beta = STC + k_{\beta=0.05}\, u_c(STC)$ | (9.9) |
| Authorized substances | | | |
| 8 | Calibration curve method according to ISO 11843-2 | Equation (9.4) with $\beta = 0.05$ | |
| 9 | Measurement uncertainty method | $CC\beta = LMR * 1.5 + k_{\beta=0.05}\, u_c(STC)$ | (9.10) |

Source: Adapted from Commission Implementing Regulation [10].

a *guard band* calculated from the MU, subtracted or added as appropriate to penalize the measurement. The higher the MU, the wider the guard band and the more challenging it is to reach the specification limit.

In this chapter and Table 9.2, it is called the MU method. The nature of the specification limit is the starting point of the calculation. As already mentioned, European regulation recognizes three types of *SL*:

| | Regulatory specification limit | Domain of application |
|---|---|---|
| *MRL* | Maximum residue limit | Many "pharmacologically active" substances are authorized [11] |
| *MC* | Maximum content | Authorized coccidiostats that contaminate other foods[12] |
| *RV* | Reference value | Prohibited substances defined in [9] |

In practice, the regulatory specfication limits are established by expert groups summoned by the European Commission. Other non-European expert groups may sometimes set different limits, for

the US-FDA [5]. For a molecule of interest, the specification limit depends on its toxicity, the foodstuff in which it is found, and its consumption level. All these factors condition the potential risks for consumer exposure and will make it possible to establish health standards.

They are supposed to be scientifically based but can sometimes be commercially or health-oriented. The reference value $RV$, defined here, is not obtained in the same way as the reference value used for interpreting clinical biology analyses. Section 8.1.2 describes the elaboration methodology for each context: food hygiene and medical diagnosis. In addition to the specification limits, other parameters are involved in the calculation. They are summarized below:

| | Parameter | Definition |
|---|---|---|
| *STC* | Screening target concentration | The concentration at or below the $CC\beta$ at which a screening action classifies the sample as potentially non-compliant and "screen positive" triggers a confirmatory test. It occurs only for prohibited substances. |
| *LCL* | Lowest calibration level | Defined by the analyst, this is the lowest concentration to which the measurement system has been calibrated. |

There is some confusion with these definitions because the official documents do not always give explicit target values. For the analyst, choosing these concentration values is a personal matter. For the screening target concentration (*STC*), it looks like an expert LOD and the lowest calibration level (*LCL*) is experimentally set during calibration.

Table 9.2 is an attempt to present in a condensed manner the various procedures and formulas for calculating $CC\alpha$ and $CC\beta$. A distinction is made between substances that are permitted and those that are not. The numbering reported in the first column of Table 9.2 is not part of the official document. It is added to better identify the diverse calculation procedures. Procedure $n°2$, although questionable, has been retained for compatibility between the old and new values, but it is intended to disappear in the short term. It is important to emphasize that all procedures allow us to calculate the results directly in the same units as the concentration.

### 9.3.4 Example of Calculation

An example of the application of the "calibration curve method according to ISO 11843-2" is provided by the Resource P worksheet. It is an application of procedures numbered 4 and 8 in Table 9.2. The starting point is to run a set of data used to construct a calibration line which must include $I > 3$ calibrator levels for which $J > 2$ replicates. These data occupy cells `B3` to `C10`.

The application of the Excel built-in function `LINEST` allows us to easily obtain the main statistics of the calibration line. In the example, the *MRL* is 200 mg/kg and is shown in cell `B11`. Finally, the results are $CC\alpha = 215$ and $CC\beta = 231$ mg/kg.

A drawback is that these criteria seem to be available only if the calibration model is linear and the data are normally distributed. So, the ISO 11843 series of standards has been expanded over the years to include other parts on various calibration models or probability distributions.

Under the 2021 EU regulation, the coverage factors $k_1$ and $k_2$ that appear in various formulas should depend on the number of measures. We are then facing two situations.

– If the number of measurements is important, the quantile of the standardized normal distribution, denoted $z$, can be used instead of the Student's $t$:

$$k_1 = k_{\alpha=0.01} = k_{\beta=0.01} = z_{(0.99)} = 2.33 \text{ (unilateral at 99\%)}$$

$$k_2 = k_{\alpha=0.05} = k_{\beta=0.05} = z_{(0.95)} = 1.64 \text{ (unilateral at 95\%)}$$

Using the built-in statistical functions of Excel, it gives:

$k_1$=NORMAL.STANDARD.INVERSE(0.95)
$k_2$=NORMAL.STANDARD.INVERSE(0.99)

– Otherwise, especially for authorized substances, the Student's t distribution law can be reasonably applied instead, taking into account the number of results collected during the validation of the method. In this case, the number of degrees of freedom or the number of effective measurements must be used $N_E$.

**Resource P Decision limit–calibration curve procedure of ISO 11843-2.**

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Resource P: Calibration curve procedure of ISO 11843-2** | | | |
| 2 | | Concentration X | Response Y | |
| 3 | | 75 | 0.583 | |
| 4 | | 75 | 0.617 | |
| 5 | | 150 | 1.150 | |
| 6 | | 150 | 1.292 | |
| 7 | | 225 | 1.933 | |
| 8 | | 225 | 1.825 | |
| 9 | | 300 | 2.375 | |
| 10 | | 300 | 2.600 | |
| 11 | | | | |
| 12 | Maximum Residue Limit | 200 | mg/kg | |
| 13 | | | | |
| 14 | Calibration curve | a1 | a0 | |
| 15 | Coefficients | 0.0084 | –0.0333 | {=LINEST(C3:C10;B3:B10;TRUE;TRUE)} |
| 16 | Standard deviations | 0.0004 | 0.0730 | {=LINEST(C3:C10;B3:B10;TRUE;TRUE)} |
| 17 | Residual std. dev. | 98.9% | 0.08428 | {=LINEST(C3:C10;B3:B10;TRUE;TRUE)} |
| 18 | Degreese of freedom | 562.52 | 6 | {=LINEST(C3:C10;B3:B10;TRUE;TRUE)} |
| 19 | | 4.00 | 0.0426 | {=LINEST(C3:C10;B3:B10;TRUE;TRUE)} |
| 20 | Alpha risk | 0.05 | | |
| 21 | Beta risk | 0.05 | | |
| 22 | kalpha | 1.943 | | =TINV(2*B20;C18) |
| 23 | kbêta | 1.943 | | =TINV(2*B21;C18) |
| 24 | Average X | 187.5 | | =AVERAGE(B3:B10) |
| 25 | SS(X) | 56250 | | =DEVSQ(B3:B10) |
| 26 | Number of replicates (J) | 2 | | |
| 27 | Number of calibrants (I) | 4 | | |
| 28 | Intermediate quantity | 7.923 | | =(C17/B15)*SQRT((1/B26)+(1/(B27*B26)) +((B12-B24)^2)/$B25) |
| 29 | Computation of CC alpha and CC Béta | | | |
| 30 | Cca | 215 | | =B12+B22*B28 |
| 31 | CCb | 231 | | =B12+2*B23*B28 |
| 32 | | | | |
| 33 | Critical response | 0.150 | | =C15+B22*C17*SQRT((1/B26)+(1/(B27*B26 ))+($B24*$B24/$B25)) |

Figure 9.3 illustrates the data of example of the Resource P worksheet and the calculation method used since 2002. It shows the relative positions of the *MRL*, *CCα*

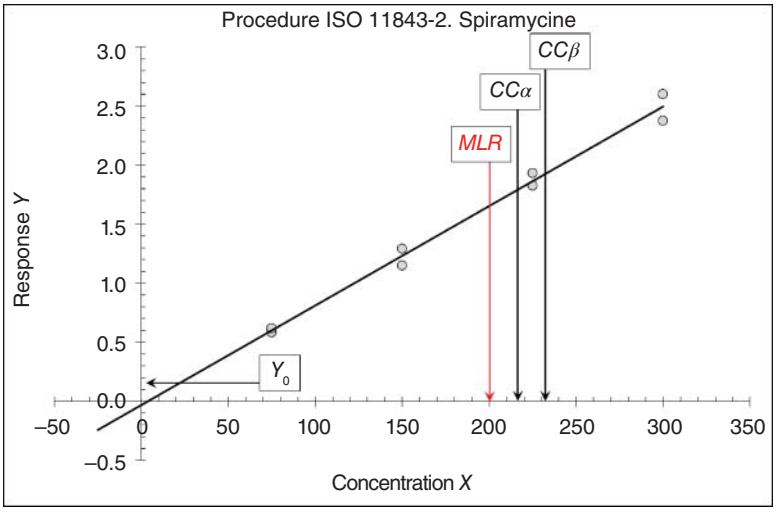



**Figure 9.3** Detection capacity for the calibration curve method according to ISO 11843-2.

and *CCβ*. As a supplement, the critical response value $Y_0$, as calculated in cell `B31` of the worksheet, is also plotted on this graph.

To calculate the combined standard uncertainty $u_c$ of the different criteria *MRL*, *MC* and *STC*, the European regulation recommends using the intermediate precision, sometimes incorrectly called intra-laboratory reproducibility as already indicated. This recommendation fully agrees with our recommendations developed in the previous chapters about MAP and uncertainty functions.

These new proposals converge on the idea that the MAP, complemented by the calculation of the uncertainty function, preferably in the form of the standard uncertainty, is perfectly suited to estimating the detection capability of methods as introduced by European legislation. In conclusion, the calculation of *CCα* and *CCβ* from an accuracy profile is a straightforward solution. No example is supplied here but any application can easily be done when considéring examples in previous chapters.

## References

1 Raposo, F. and Ibelli-Bianco, C. (2020). Performance parameters for analytical method validation: controversies and discrepancies among numerous guidelines. *Trends in Analytical Chemistry* 129: 115913.

2 Currie, L.A. (1999). Detection and quantification limits: origins and historical overview. *Analytica Chimica Acta* 39: 127–134.

3 ICH (2022). International Council for Harmonisation of technical requirements for pharmaceuticals for human use (ICH) Guideline Q2(R2) on validation of analytical proceduresStep 2b, Amsterdam.

4 Russ, C.W. IV, Prest, H., and Wells, G. (2011). Why use signal-to-noise as a measure of MS performance when it is often meaningless? *Spectroscopy* https://www.spectroscopyonline.com/view/why-use-signal-noise-measure-ms-performance-when-it-often-meaningless (accessed 5 September 2023).
5 Food and Drug Administration (FDA) (2018). *Bioanalytical Method Validation Guidance for Industry*. Washington, DC: Office of Communications, Division of Drug Information Center for Drug Evaluation and Research.
6 Council Directive 96/23/EC of 29 April 1996 on measures to monitor certain substances and residues thereof in live animals and animal products and repealing Directives 85/358/EEC and 86/469/EEC and Decisions 89/187/EEC and 91/664/EEC.
7 2000/657/EC: Commission Decision of 16 October 2000 adopting Community import decisions for certain chemicals pursuant to Council Regulation (EEC) No 2455/92 concerning the export and import of certain dangerous chemicals.
8 Standard ISO 11843-2 (2000). *Capability of Detection — Part 2: Methodology in the Linear Calibration Case*. Genève: ISO.
9 Commission Regulation (EU) 2019/1871 of 7 November 2019 on reference points for action for non-allowed pharmacologically active substances present in food of animal origin and repealing Decision 2005/34/EC.
10 Commission Implementing Regulation (EU) 2021/808 of 22 March 2021 on the performance of analytical methods for residues of pharmacologically active substances used in food-producing animals and on the interpretation of results as well as on the methods to be used for sampling and repealing Decisions 2002/657/EC and 98/179/EC.
11 Commission Regulation (EU) No 37/2010 of 22 December 2009 on pharmacologically active substances and their classification regarding maximum residue limits in foodstuffs of animal origin.
12 Commission Regulation (EC) No 124/2009 of 10 February 2009 setting maximum levels for the presence of coccidiostats or histomonostats in food resulting from the unavoidable carry-over of these substances in non-target feed.

# 10

# Examples of MU Application

## 10.1 Standard Addition Method and Drug Quality

Access to appropriate, often patent-protected medecines, can be extremely difficult for low- and middle-income countries, and, in some cases, for high-income countries. For the latter, vulnerable populations, such as prisoners, drug addicts, or migrants, who have limited access to certain therapies, could be affected. In addition to its ethical aspects, this issue raises public health problems for controlling certain pandemics, such as acquired immuno deficiency syndrome (AIDS).

The price of original patented medicines appears to be a limiting factor for widespread treatment. Indeed, many indigent patients are treated by public health institutions whose resources are limited and which therefore have difficulty fulfilling their missions. For this reason, in several countries, including the United States, Canada, and European nations such as Switzerland, nonprofit organizations known as buyers' clubs have been set up to facilitate access to these essential medicines for vulnerable populations.

These buyers' clubs use a "right to import unlicensed ready-to-use medicines from individuals" defined in international agreements [1]. Even if buyers' clubs select pre-qualified manufacturers for import, the possibility of receiving drug products of insufficient quality cannot be excluded due to the high frequency of falsified molecules coming from exporting countries.

Therefore, a quality control test upon arrival of the drug could help to ensure patient safety. However, this control must be carried out, differently from the classical control established in the pharmaceutical industries, because there are two main constraints to developing a satisfactory assay method:

- The first is the lack of knowledge of the exact nominal content of the drug and the impossibility of having validation materials with guaranteed content.
- The second is the possible presence of matrix interferences that can bias the measurements.

In addition, external calibration for quantification is made difficult by the small number of samples available and the large potential differences in composition for a drug containing the same active principal ingredient (API). To overcome these obstacles, the control of various antiretroviral drugs received by buyers' clubs could

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

be based on standard addition method (SAM). The pros and cons of this quantification mode are extensively described in Section 1.4.1.

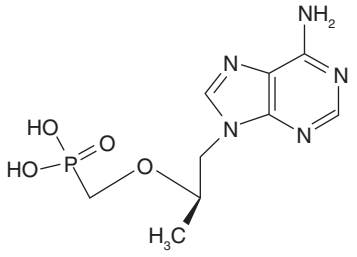
It is to be remembered that SAM is generally said to avoid matrix effects and simultaneously corrects calibration biases due to diverse sources, such as medicine coating, excipients, or interfering matrix components. It is also extremely useful when it is not possible to have surrogate validation materials with content which is exactly known.

There are various SAM operating procedures, and the chosen one consists in adding increasing amounts of a standard solution of the analyte to be determined in constant aliquots of the sample. In this case, the concentrations of all the matrix components are kept constant, except for the molecule to be determined, here being the API. Nevertheless, since the sample predicted concentration is calculated by extrapolating a response function, this method may induce a bias – generally small – due to possible multiplicative effect interferences.

The procedure, although very efficient in many cases, is often criticized from a statistical point of view because extrapolation is prone to increase prediction uncertainty, especially when the amount of spike is not properly adjusted with respect to the initial signal measured in the tested sample.

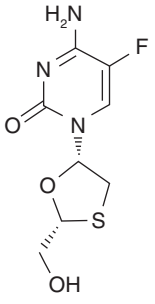
When it is possible to make several simultaneous measurements, typically at different wavelengths, the technique called H-point standard addition method (HPSAM) dramatically reduces the risk of not detecting interferences. In the original version developed for a spectrophotometry_based method, it consisted of multiple determinations of an analyte by SAM at several wavelengths corresponding as much as possible to different signals varying with the analyte concentration and constant for the interferent [2]. It is then possible to simultaneously compute several in-sample calibration curves.

Although they have different slopes because the sensitivity may vary, in the absence of important additive bias due to matrix effects, they must intersect with the concentration axis at close points and give close extrapolated concentrations. If the additive bias is critical, a modified HPSAM was proposed, including chemical modifiers [3].




Tenofovir (TDF)

The purpose of this example is to show how to estimate the measurement uncertainty (MU) when applying the H-point operating procedure. It consists in the determination of two antiretroviral active ingredients present in various pharmaceutical formulations used for the treatment of AIDS, namely tenofovir (TDF) and emtricitabine (FTC).




Emtricitabine (FTC)

Two approaches have been used to estimate the MU. The first one uses the classical statistical information available from the extrapolated concentration obtained by SAM, the second one is based on a complete experimental design to build an accuracy profile to demonstrate the continuity between the MAP procedure and the estimation of MU, whatever the quantification method.

It should be noted that in many similar cases, complete validation is impractical or even impossible. This type of procedure requires relatively massive quantities of validation materials of exactly known contents prepared on purpose. Moreover, in the case of drugs imported by buyers' clubs, the diversity of formulations would require specific validation for each specialty.

Capillary zone electrophoresis (CZE) is an analytical technique recognized as performing well for the control of adulterated pharmaceutical formulations and/or the quality control of drugs, in general. This technique is, therefore, particularly well suited to the case in hand: it is economical thanks to a reduced solvent consumption of the order of 1 µl per analysis, and it avoids the use of expensive and environmentally damaging organic solvents.

Quantification by CZE is relatively easy, and many studies have shown that its performance is similar to conventional methods, such as liquid chromatography. In addition, for this study, the instrument is coupled to a diode array UV detector (DAD), which allows the measurement of responses at different wavelengths, as required for the HPSAM, in this case, 200, 210, and 254 nm.

The preparation of the sample is simple: it is ground, then suspended in a solvent, which after dilution, forms the injected solution. For example for a drug advertised at 240 mg of active ingredient per tablet, the dilution factor applied was 1 : 2500 so that it to be around 0.1 mg/ml in the final solution and be in a range where the UV signal is in the linear range of the detector used.

### 10.1.1 SAM Without Replication

According to the chosen SAM procedure, equal volumes of the sample solution are taken, and known quantities of a standard solution are added separately. In this case, in addition to the original sample, two spikes are done, which correspond to approximately 50% and 100% of the expected signal for the likely nominal content. This is a standard protocol. All these aliquots are then brought to the same final volume and measured by CZE.

More details on this procedure, the reagents used, and the instrumental conditions are available in [4]. In the absence of significant matrix effects, at the chosen

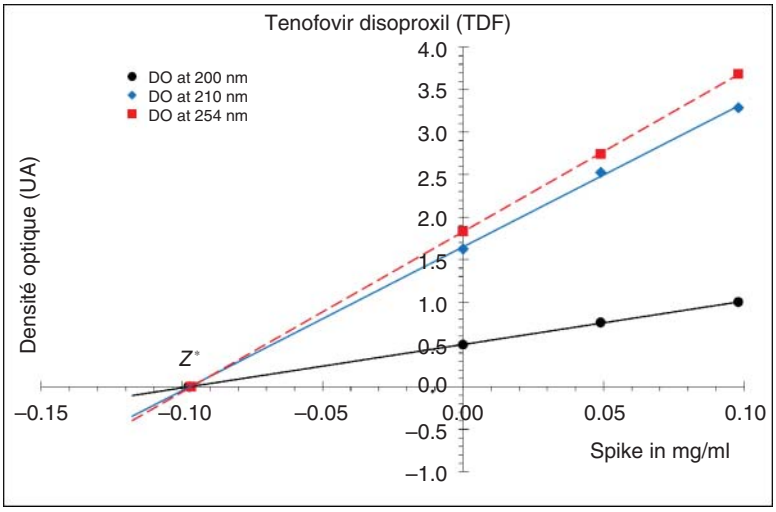



**Figure 10.1** Example of assay obtained by SAM on a pharmaceutical product containing TDF. Three estimates of extrapolated concentration can be computed. For sake of clarity only one value is shown.

wavelengths, the three calibration lines, although with different slopes, intersect with the concentration axis for extrapolated concentrations remarkably close to each other. This approach makes an optimized use of the effort required for sample preparation allowing several replicates to be obtained for the same sample.

Figure 10.1 shows an example of an assay performed under the described conditions for a medicine containing TDF at the claimed nominal content of 250 mg/tablet. Measurements were made at three wavelengths, namely 200, 210, and 254 nm, with standard additions at 0.049 and 0.098 mg/ml. As usual, the signal is shown on the vertical axis, expressed in arbitrary units (AU), and the amounts of analyte added on the horizontal axis.

The estimated content in the working sample denoted $Z^*$(star) is obtained by extrapolating the regression lines to the point where the response is 0. As shown in the figure, it is in the negative part, around 0.1 mg/ml. Since a dilution factor of 1 : 2500 was applied, the concentration per tablet is around 250 mg. For each wavelength, three spiked samples are measured.

Let us note each wavelength with the index $j$, with $1 \leq j \leq J$ since the measurements, once converted into concentration, will be assimilated to replicates. Each regression line connecting the measurement values is calculated using the same equations and notations as in Section 2.3.1, except for subscripts.

| | |
|---|---|
| $X_n$ | Spike concentration in mg/ml |
| $1 \leq n \leq N$ | Number of spiked points, in the example $N = 3$ |
| $Y_n$ | Optical density measured in AU |
| $Z^*$ | Extrapolated sample concentration |

Calibration model

$$Y_n = a_0 + a_1 X_n + E_n \tag{10.1}$$

Extrapolated concentration

$$Z^* = \frac{a_0}{a_1} \tag{10.2}$$

Variance of the extrapolated concentration

$$s^2(Z^*) = \left(\frac{s_E}{a_1}\right)^2 \times \left[\frac{1}{N} + \frac{\overline{Y^2}}{a_1^2 \sum_n (X_n - \overline{X})^2}\right] \tag{10.3}$$

Residual standard deviation of the regression

$$s_E = \sqrt{\frac{\sum_n (Y_n - \widehat{Y_n})^2}{N-2}} \tag{10.4}$$

The calculation is repeated at each wavelength, and finally, three extrapolated measurement values, denoted $Z_j^*$ with $1 \le j \le J$ are collected. Each corresponds to three measurements: one without spikes and two with different spikes. These measurements will be assumed to be replicates under repeatability conditions, as they cannot be considered as obtained under intermediate precision conditions since the measuring techniques are not identical, since different wavelengths are used.

The regression line coefficients are estimated by OLS regression. The three extrapolated concentrations must be multiplied by the dilution factor of 1: 2500 to recover the expected contents in the tablet. For each extrapolated concentration, it is also possible to calculate a variance using the formula (10.3) proposed by [5]. The number of associated degrees of freedom is then ridiculously small and equal to 1. The variance is used as an estimate of the standard variance of an extrapolated result $s^2(Z^*) = u^2(Z^*)$.

The implementation of the formulas in Excel® is simple when using the built-in matrix function `LINEST` because it directly provides the residual standard deviation $s_E$ as explained in Section 2.3.1 and shown on the Resource Q worksheet. To use `LINEST` in its matrix form, please refer to the Excel® user's manual.

For the illustrated application of Resource Q, the calculation for the SAM is done for a single wavelength at 200 nm. The results returned by `LINEST` are surrounded

**Table 10.1** TDF measurements in a drug declared at 245 mg/tablet, obtained by HPSAM at different wavelengths ($1 \le j \le 3$).

| Wavelengths (nm) | Extrapolated content (mg/ml) | Variance | Level found (mg/tablet) | Recovery rate (%) |
|---|---|---|---|---|
| 200 | 0.0977 | $1.111 \times 10^{-5}$ | 244.1 | 99.6 |
| 210 | 0.0973 | $5.567 \times 10^{-5}$ | 243.2 | 99.3 |
| 240 | 0.0967 | $2.303 \times 10^{-6}$ | 241.8 | 98.7 |
| Average $\overline{Z}$ | 0.0972 | | | 99.2 |

by a light line, and the residual standard deviation $s_E = 0.00777$ by a thick border in cell C13. The most complicated formula is the Intermediate quantity in cell B22. It corresponds to the quantity between square brackets in Eq. (10.3).

**Resource Q Calculation of the SAM extrapolated concentration (Excel).**

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Resource Q: Calculation of the extrapolated concentration by SAM** | | | |
| 2 | Tenofovir disoproxil (TDF) | | | |
| 3 | Dilution factor | 2500 | | |
| 4 | Nominal value (mg/tablet) | 245 | | |
| 5 | | | Optical density (OD) | |
| 6 | | Spike in mg/ml | DO at 200 nm | |
| 7 | | 0.000 | 0.50 | |
| 8 | | 0.049 | 0.76 | |
| 9 | | 0.098 | 1.00 | |
| 10 | | | | |
| 11 | SAM calibration line | Slope a1 | Intercept a0 | Formulas |
| 12 | | 5.115 | 0.499 | {=LINEST(C7:C9;B7:B9;1;1)} |
| 13 | | 0.1121 | 0.0071 | {=LINEST(C7:C9;B7:B9;1;1)} |
| 14 | | 0.9995 | 0.00777 | {=LINEST(C7:C9;B7:B9;1;1)} |
| 15 | | 2080.6 | 1.000 | {=LINEST(C7:C9;B7:B9;1;1)} |
| 16 | | 1.26E-01 | 6.04E-05 | {=LINEST(C7:C9;B7:B9;1;1)} |
| 17 | | | | |
| 18 | Extrapolated concentration (Z*) | 0.0977 | =C12/B12 | |
| 19 | Average Z | 0.7501 | =AVERAGE(C7:C9) | |
| 20 | Number of spikes (N) | 3 | =COUNTA(B7:B9) | |
| 21 | se/a1 | 0.001519 | =C14/B12 | |
| 22 | 1/N | 0.3333 | =1/B20 | |
| 23 | Intermediate quantity | 4.8123 | =B22+((B19*B19)/(B12*B12*DEVSQ(B7:B9))) | |
| 24 | s(Z*) | 3.3327E-03 | =B21*SQRT(B23) | |
| 25 | $s^2$(Z*) | 1.1107E-05 | =B24^2 | |
| 26 | Recovered conc. (mg/tablet) | 244.1 | =B18*B3 | |
| 27 | Recovery yield | 99.6% | =B26/B4 | |
| 28 | | | | |

The same worksheet is copied and updated for each wavelength. The final results for the three wavelengths are gathered in Table 10.1. Recovery rates are obtained by accounting for the dilution factor applied to the extrapolated contents compared to the content announced by the manufacturer.

Finally, $J = 3$ replicate results are combined, each with an estimated standard variance, i.e. squared standard uncertainty. The addition of covariances is probably necessary, but it is impossible to calculate them, given the small number of degrees of freedom. Since each measurement value is considered one replicate, the analytical result can be expressed as an average. By making this approximation, it is also possible to estimate the repeatability variance by taking the average of the variances of the replicates (formula 10.5).

The resulting situation is comparable to the one described in Section 8.4.3 about calculating the MU of a result obtained by averaging replicates achieved under repeatability conditions. This gives the following formulas:
Mean extrapolated concentration

$$\overline{Z} = \frac{\sum_{j=1}^{J} Z_j^*}{J} \tag{10.5}$$

Repeatability variance of the mean

$$s_r^2 = \frac{\sum_{j=1}^{J} s^2\left(Z_j^*\right)}{J} \tag{10.6}$$

Standard variance of the mean content

$$u^2(\overline{Z}) = \frac{s_r^2}{J} = \frac{\sum_{j=1}^{J} s^2\left(Z_j^*\right)}{J}/J \tag{10.7}$$

Numerical application to the results for the TDF analysis gives:

| Parameters | Extrapolated value | Predicted concentration | Recovery yield |
|---|---|---|---|
| Unit | mg/ml | mg/tablet | % |
| $J$ | 3 | | |
| $\overline{Z}$ | 0.0972 | 243.0 | 97.22 |
| $s_r^2$ | $2.303 \times 10^{-5}$ | | |
| $u^2(\overline{Z})$ | $7.676 \times 10^{-6}$ | | |
| $U(\overline{Z})$ | 0.00554 | 13.9 | |
| $UR\%(\overline{Z})$ | 5.7% | 5.7% | |
| Coverage interval | 0.0917 | 229.2 | 91.68 |
| | 0.1028 | 256.9 | 102.76 |

Assuming that the variance in formula (10.7) is the standard variance may be abusive, insofar as it was obtained under repeatability conditions. But the small number of degrees of freedom is a limitation that impairs the quality of this estimate. Nevertheless, it represents a first estimate of the MU that is easily computed. The relative uncertainty thus obtained, close to 6%, is relatively high if one considers the ±5% acceptance criterion classically adopted in pharmacy for active principles but it can be considered operative in the present context.

Figure 10.2 shows a similar study performed on a drug containing FTC. In this case, the standard additions are slightly different (0.040 and 0.080 mg/ml) to fit the expected nominal content, declared at 200 mg/tablet.

The dilution factor is the same. The results obtained from these data are as follows:

| Symbol | Extrapolated FTC content (mg/ml) | Level found (mg/tablet) | Recovery yield (%) |
|---|---|---|---|
| $J$ | 3 | | |
| $\overline{Z}$ | 0.0799 | 199.9 | 99.94 |
| $s_r^2$ | $3.612 \times 10^{-5}$ | | |
| $u^2(\overline{Z})$ | $1.204 \times 10^{-5}$ | | |
| $U(\overline{Z})$ | 0.00694 | | |
| $UR\%(\overline{Z})$ | 8.7% | | |
| Coverage interval | 0.0730 | 182.5 | 91.26 |
| | 0.0869 | 217.2 | 108.61 |

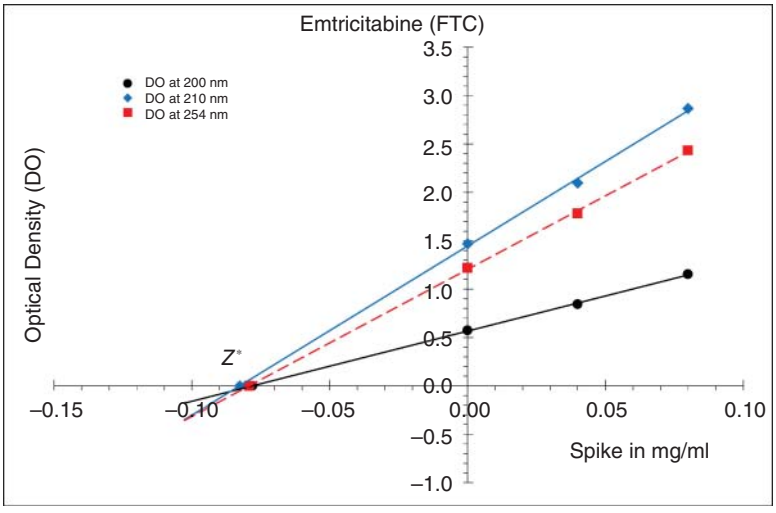



**Figure 10.2** Determination of FTC, by standard additions to a pharmaceutical product announced at 200 mg/tablet.

In this example for FTC, the analytical method achieves an inferior performance since the estimated relative uncertainty is around 9%. The proposed H-point procedure and calculation method were applied to six imported pharmaceutical preparations containing TDF and FTC. Estimated relative uncertainties range from 4 to 19%. In each case, the inverse-predicted mean content is calculated from 9 absorbance measurements, as well as its coverage interval.

Figure 10.3 summarizes these results. On the same graph, the ±5% acceptance limits classically used in the pharmaceutical field for formulations are plotted in

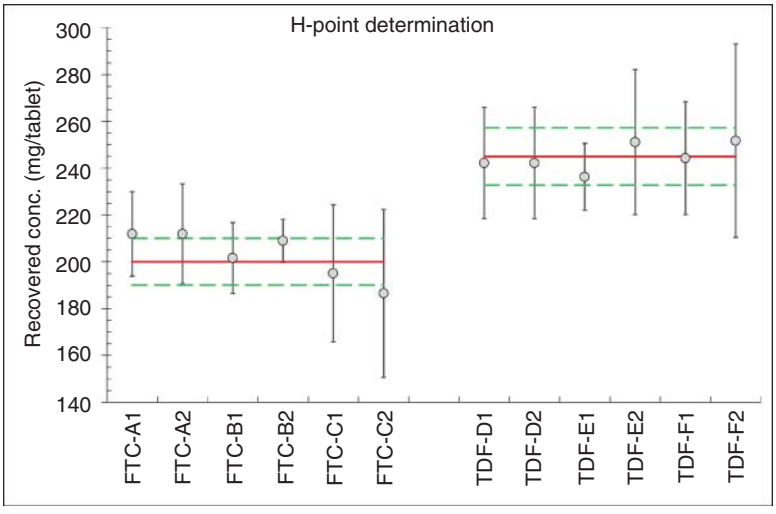



**Figure 10.3** Average levels found (mg/tablet) with the coverage intervals for six medicines noted from A to F. Expected nominal content is represented by broad red lines surrounded by the acceptance interval of ±5% as green dotted lines.

green dotted lines. It appears that for several controls, such as FTC-C2, the situation is quite degraded.

### 10.1.2 SAM with Replication

In Figure 10.3, the full replication number of the analysis is indicated by the last digit of the label. Finally, each of the six samples, three containing FTC and three with TDF, were independently duplicated. Thus, the experimental design includes several series of measurements. It is then possible to extract a between-series variance and calculate the uncertainty of the inverse-predicted concentration in closer accordance with the GUM recommendations.

As stated in Section 8.4.3, it is not an optimal experimental design since there are only two series of three replicates, while it is recommended to make three series of two replicates. The notations used for the experimental design are as follows:

| | |
|---|---|
| $I$ | Number of series $I = 2$ |
| $J$ | Number of replicates per series $J = 3$ |
| $Z^*_{i,j}$ | Extrapolated content (at a wavelength $q$) for a series $I$ |
| $\overline{Z^*_i}$ | The average content of series $I$ (Eq. 10.5) |
| $s^2\left(Z^*_{i,j}\right)$ | The variance of an extrapolated content (Eq. 10.6) |
| $T$ | Inverse-predicted concentration in a tablet, $T = \overline{\overline{Z}} \times \frac{1}{\text{Dilution Factor}}$ |

This occurs to the following formulas:

Grand mean

$$\overline{\overline{Z}} = \frac{\sum_{i=1}^{I} \overline{Z^*_i}}{I} \tag{10.8}$$

Repeatability variance

$$s_r^2 = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} s^2(Z)^*_{i,j}}{IJ} \tag{10.9}$$

Between-series variance

$$s_B^2 = \frac{\sum_{i=1}^{I} \left(\overline{Z^*_i} - \overline{\overline{Z}}\right)^2}{I-1} \tag{10.10}$$

Standard variance of the grand mean

$$u^2(\overline{\overline{Z}}) = \frac{s_B^2}{I} + \frac{s_r^2}{IJ} \tag{10.11}$$

Numerical application of these formulas to six batches of drugs labeled A to F from two manufacturers gave the following results, grouped in Table 10.2.

Figure 10.4 illustrates these results. In general, the inverse-predicted content is obtained with an almost acceptable uncertainty if one considers the classic acceptance range of the pharmaceutical field for this determination.

**Table 10.2** Calculation of measurement uncertainty for different batches.

| Sample | Average $\overline{Z_1^*}$ | Average $\overline{Z_2^*}$ | $s_r^2$ $\times 10^{-5}$ | $s_B^2$ $\times 10^{-5}$ | $u^2(\overline{\overline{Z}})$ $\times 10^{-5}$ | $U(\overline{\overline{Z}})$ $\times 10^{-3}$ | *UR* % (%) |
|---|---|---|---|---|---|---|---|
| FTC-A | 0.083 | 0.084 | 1.529 | 0.055 | 1.047 | 6.5 | 7.7 |
| FTC-B | 0.082 | 0.083 | 0.639 | 0.044 | 0.448 | 4.2 | 5.1 |
| FTC-C | 0.080 | 0.078 | 4.643 | 0.140 | 3.165 | 11.3 | 14.4 |
| TDF-D | 0.098 | 0.098 | 2.329 | 0.000 | 1.552 | 7.9 | 8.0 |
| TDF-E | 0.094 | 0.100 | 2.294 | 1.540 | 2.299 | 9.6 | 9.6 |
| TDF-F | 0.097 | 0.100 | 4.522 | 0.392 | 3.210 | 11.3 | 11.3 |

| Sample | Predicted concentration *T* (mg/tablet) | *U*(*T*) | Coverage interval of *T* | |
|---|---|---|---|---|
| FTC-A | 209.6 | 16.1 | 194 | 226 |
| FTC-B | 205.5 | 10.5 | 195 | 216 |
| FTC-C | 197.8 | 28.4 | 169 | 226 |
| TDF-D | 245.6 | 19.7 | 226 | 265 |
| TDF-E | 242.6 | 23.3 | 219 | 266 |
| TDF-F | 246.5 | 27.9 | 219 | 274 |

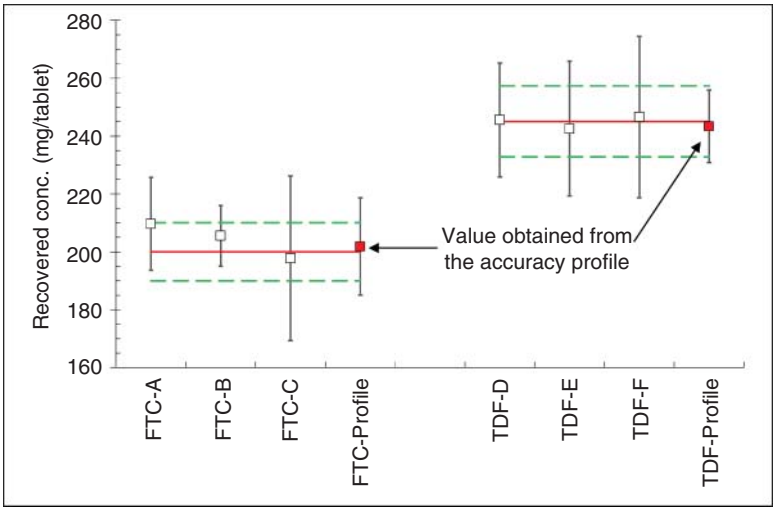



**Figure 10.4** MU estimates of six drug lots for two nominal strengths of 200 and 245 mg/tablet. The error bars represent the coverage intervals. On the same graph, the accuracy profile results are plotted as a red solid squares. Same legend as Figure 10.3.

The control of TDF (tenofovir disoproxil) thus appears to be somewhat more effective than that of FTC (emtricitabine). Considering the measurements as two series and taking into account the between-series variance appears to significantly reduce the MU on the average predicted content. The new values are obtained by combining 12 absorbance measurements (instead of 9), which is relatively small. In Figure 10.4 the results obtained by applying the procedure derived from the accuracy profile are also reported as a black square, labeled FTC-Profile and TDF-Profile. The following chapter presents other possible data processings.

### 10.1.3 Estimation from Method Accuracy Profile

To confirm the usefulness of MU estimates obtained with HPSAM, it was decided to implement an experimental design corresponding to that recommended for obtaining a MAP using synthetic samples of known concentration. But such validation materials with known assigned concentrations were not available; therefore, a validation material was prepared by crushing and mixing 20 tablets of the drug, said to contain a mixture of 245 mg TDF and 200 mg FTC per tablet.

From the homogenized batch, independent determinations of the two APIs were performed with the classical SAM procedure, i.e. by making two additions of the same stock solution, as described above. In this case, the measurements were performed at a single wavelength of 200 nm. The experimental design itself consisted of three series of six replicates per series. Table 10.3 summarizes this dataset called ARV for antiretroviral drugs, while Table 10.4 indicates the inverse-predicted concentrations of TDF and FTC.

**Table 10.3** ARV – description of the ARV dataset.

| Title | ARV (antiretroviral drugs) |
|---|---|
| Reference article | [4] |
| Measurand | Percentage recovery of active ingredient, expressed as a function of the expected nominal content |
| Method | Zone capillary electrophoresis with UV detection at multiple wavelengths. |
| Acceptance interval | ±5%. Value classically used for fraudulent drug analysis |
| Validation materials | Mixture of 20 crushed tablets |
| Validation plan | Number of series ($I = 3$); number of replicates ($J = 6$/series); number of levels or products ($K = 2$) |
| Calibration plan | See the reference article. |
| Total number of measures | 18 measurements on each of the API present, tenofovir (TDF) and emtricitabine (FTC) in a surrogate sample. Each of these measurements is made at a single wavelength with 2 standard additions and a zero addition, for a total of 108 measurements. |
| Inverse-predicted concentrations | Table 10.4 |

**Table 10.4** ARV – experimental design and levels found for the two molecules.

| | Tenofovir disoproxil (TDF) 245 mg/tablet | | | Emtricitabine (FTC) 200 mg/tablet | | |
|---|---|---|---|---|---|---|
| **Replicates** | **Series 1** | **Series 2** | **Series 3** | **Series 1** | **Series 2** | **Series 3** |
| 1 | 239.1 | 251.9 | 245.7 | 214.9 | 205.7 | 194.9 |
| 2 | 243.5 | 244.9 | 233.7 | 191.2 | 201.6 | 209.3 |
| 3 | 240.5 | 245.8 | 234.5 | 202.4 | 203.9 | 199.5 |
| 4 | 250.4 | 244.7 | 239.3 | 215.7 | 192.0 | 213.5 |
| 5 | 244.5 | 239.6 | 242.8 | 192.3 | 194.8 | 200.0 |
| 6 | 250.4 | 251.6 | 237.3 | 205.7 | 190.5 | 205.2 |

**Table 10.5** ARV – estimated MU and coverage interval from MAP.

| **Parameter** | **Symbol** | **TDF** | **FTC** |
|---|---|---|---|
| Average predicted content | $T$ | 243.3 | 201.8 |
| Repeatability standard deviation of | $s_r$ | 4.7262 | 8.2027 |
| Between-series standard deviation | $s_B$ | 3.4513 | 0.0 |
| Intermediate precision std. dev. | $s_{FI}$ | 5.8522 | 8.2027 |
| Effective measures | $N_E$ | 8.07 | 16.62 |
| Coverage probability | $\beta$ | 95% | 95% |
| Coverage factor | $k_{IT}$ | 2.30 | 2.11 |
| Coverage interval | | 228.9 | 184.0 |
| | | 257.8 | 219.7 |
| Expanded uncertainty | $U(T)$ | 12.6 | 16.9 |
| Relative uncertainty | $UR\,\%\,(T)$ | 5.16% | 8.35% |

From these data, various validation parameters are calculated using the Resource Q worksheet. The main results are reported in Table 10.5. Unlike other accuracy profiles in this book, results are not expressed as recovery yields referring to the assigned nominal content. This mode of expression cannot be used because the tablet's content is unknown; it is only assumed.

The proportion $\beta$ applied for these calculations is 95% since the goal is to evaluate the 95% expanded uncertainty. This means that 95% of the probable values of the inverse-predicted concentrations lie in the interval [229; 258] mg/tablet for an assumed level of 245 mg of TDF and between [184; 220] at 200 mg/tablet of FTC. Because the content announced by the manufacturer of the imported drug is included within the estimated coverage interval, the product can be considered conforming. The coverage intervals depend on the relative uncertainty, which is about 5% for TDF and 8% for FTC at this concentration. The two values make it impossible to reach an acceptance interval of ±5%.

If the products are considered conform, the recovery yields apply to the nominal content of the drug products. Thus, the coverage intervals of the recovery yields are [93%; 105%] for TDF and [92%; 110%] for FTC, respectively. They can be considered satisfactory since the value of 100% is included in the coverage intervals.

This comment highlights that a decision based on a coverage interval is not exactly a statistical null hypothesis test, although the conclusion is comparable. It should be remembered that for the buyers' club, it is impossible to have validation materials with exactly known content. The value claimed by the manufacturer, is considered here as a true value. However the potential bias of the method should be corrected using the standard addition method. This does not pose any problem for estimating the MU; on the contrary, since the GUM recommends updating all significant sources of bias before estimating the MU.

For FTC, the main source of uncertainty is entirely due to repeatability with a value of 4.1%; the between-series effect (in this case, inter-day) is zero. In the case of TDF, while the MU is smaller, the repeatability and the between-series effect are approximately equivalent.

When graphically comparing the various approaches (Figures 10.3 and 10.4), the HPSAM, repeated in two series, is adequate if it is required that the acceptance interval must be within ±10%. This can easily be explained by a more significant number of degrees of freedom than when only one measurement is used. In this case, the risk of a false decision in releasing a batch of imported drugs is not negligible in the context of the buyers' clubs.

## 10.2 Method Comparison Using Uncertainty

### 10.2.1 Analyte Defined by the Operating Procedure

As previously discussed, some analytes are defined *per se*, i.e. by the analytical operating procedure. This usually means that their exact chemical formula, in terms of clearly identified molecular composition, is not fully established. There are many examples, such as blood cholesterol, moisture in various forms and matrices, dietary fiber, protein, total organic carbon, and oil acidity. It could be considered that the standard operating procedure defines the analyte.

These analytes were often introduced a long time ago with old methods and, over time, have become references for medical diagnosis, technical control, or commercial exchange. For example in Europe, the total fat content of raw milk is used for payment for quality. The AOAC and FDA have taken to referring to these as *gold standard* methods. It seems preferable to refer to them as official or reference methods, as gold has no longer been a monetary standard since 1971 under the presidency of Richard Nixon.

The classic disadvantages of traditional methods are that they are time-consuming, labor-intensive, and often require hazardous reagents; they are therefore expensive, and many publications suggest replacing them with alternative methods that are faster and cheaper.

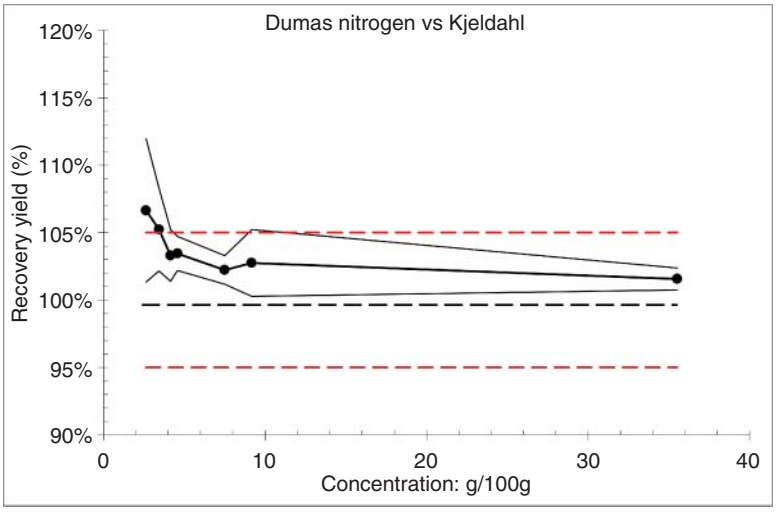



**Figure 10.5** Accuracy profile of Dumas's method applied to dairy products using Kjeldahl method as reference. $\beta$-ETI with $\beta = 80\%$. Acceptability $\pm 5\%$.

The so-called Kjeldahl nitrogen is one of the official methods for measuring total proteins in foods and feedstuffs. The original procedure was developed at the end of the nineteenth century by a Danish chemist. It consists of an acid attack of the test portion by concentrated sulfuric acid, boiling at reflux for hours. This action is supposed to destroy the molecules containing organic nitrogen (proteins and nucleic acids) and mineralize them into ammonium sulfate.

Since its invention, even if there are many variants, the principle has remained the same, and the damage due to acid-generated vapors is devastating for laboratories. Finally, the conversion of Kjeldahl nitrogen into total proteins is done by multiplying the result by a consensus factor that depends on the matrix type.

The Dumas method, on the other hand, is not much newer, but it has at least two advantages: speed and the absence of strong acid. It consists of total oxidation of nitrogen by combustion in the presence of excess oxygen. For a long time, it was difficult to apply, but today it is combined with an efficient catharometric determination of nitrogen in gaseous form after the elimination of sulfur and carbon oxides and a reduction of nitrogen oxides to gaseous nitrogen.

A few years ago, these two methods were automated by various manufacturers. For various foodstuffs, such as dairy and cereal products, the question arose whether the official Kjeldahl method could be replaced by the faster and less expensive Dumas' method. Several validation studies were carried out. Figure 10.5 presents one of these studies for dairy products using the MAP[1].

1 Unpublished personal data.

### 10.2.2 Kjeldahl and Dumas Method Comparison

The experimental design is classical, formed by $I = 3$ series with $J = 3$ replicates per series. The number of validation materials is $K = 7$. These are commercial dairy products such as yoghurt, fermented milk, or fresh cheese, except for the most concentrated material which is gelatin. These are, therefore, matrices without standard additions, which explains the absence of levels between 10 and 30 g of nitrogen/100 g, as dairy products do not naturally reach such concentrations.

The same experimental designs were applied for both Kjeldahl and Dumas methods. Since the standardized Kjeldahl method is considered the reference method capable of defining the reference value, the target content assigned to each material was established by the average of $I \times J = 9$ replicate measurement values performed with the Kjeldahl method.

The acceptance interval is $\pm 5\%$. This is narrow because manufacturers require the results of the two methods to be remarkably close for regulatory labeling reasons. To be converted into protein, the Kjeldahl nitrogen content is multiplied by 6.25: a simple difference of 0.16 g of Kjeldahl nitrogen becomes 1 g of total protein. It can create problems in the case of verification of the claimed nutrient composition.

According to the MAP of the Dumas method in Figure 10.5, a systematic overestimation of the concentration is observable when compared to the Kjeldahl reference method is observed. In absolute value, this bias varies between 0.14 and 0.25 g/100 g depending on the matrix type.

With available data, it is possible to build the MAP of the two methods and obtain their respective uncertainty functions. They are combined in Figure 10.6. They have similar shapes, the Kjeldahl method having even slightly lower performance when compared to Dumas's method. At low concentrations, the bias is between 5 and 7%, and the relative uncertainty is of the same magnitude.

Therefore, the bias becomes very bothersome when controlling low concentration foods, such as milk or yoghurt. Because the Kjeldahl method is time-consuming and polluting, a trend has been to reduce the time of mineralization in laboratories. In addition, various toxic or polluting catalysts used in the past have been replaced by others that do not have the same efficiency.

After this study, the Kjeldahl method has been suspected of not being as accurate as its *gold standard* label would suggest. The various materials used to construct the profile of the Dumas method were resampled and reanalyzed with the Kjeldahl operating procedure, but extending the duration of mineralization.

The new profile obtained with these modified reference measures is presented in Figure 10.7. The systematic bias has almost disappeared, and the method meets the requirements of the end users over the whole validation range. In this case, it is not mandatory to add to the MU the uncertainty of a correction factor: new reference values were assigned to the validation materials, but Dumas' measurements were not modified.

This example demonstrates that, when validating an alternative method against a reference method, the existence of a possible bias induced by the so-called reference method is always possible and must be carefully evaluated. This is misleading when

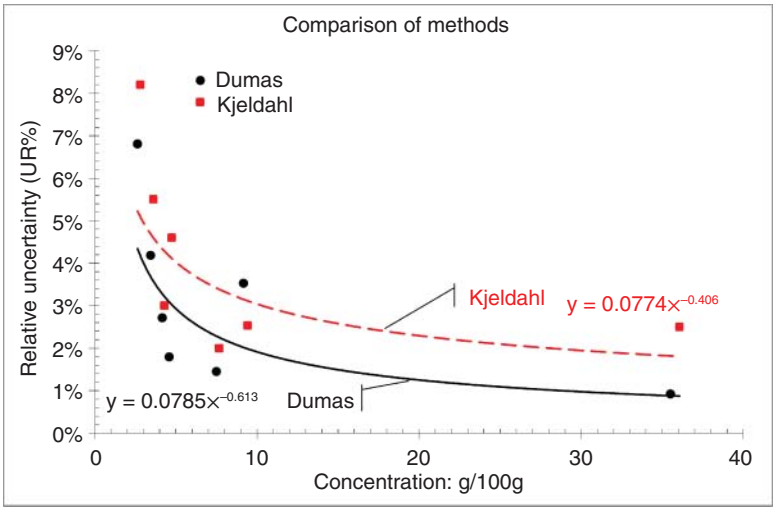



**Figure 10.6** Comparison of the uncertainty functions of both methods used to determine total nitrogen in foods.

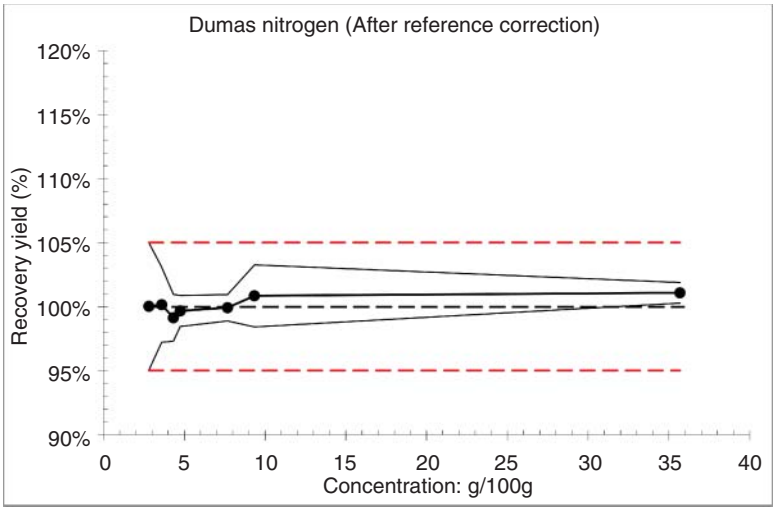



**Figure 10.7** Accuracy profile of Dumas's method applied to dairy products compared to improved Kjeldahl method.

it is a *per se* defined method. This drawback only appears when the analyte is defined by the method. To detect this additive bias, it was necessary to focus on low concentrations because the trueness was expressed as a recovery rate.

Today the Dumas method is considered accurate and represents a credible alternative to the Kjeldahl method for analyzing the protein content of many foods. Even international standards address this method, such as the standard ISO 16634-2:2016 for cereals, pulses, and milled cereal products.

To conclude on the alternative Kjeldahl methods, there are other operating procedures based on near-infrared spectrometry (NIRS), which has the advantage of being non-destructive. The comparison with the Kjeldahl reference is done by a rather special chemometric treatment, partial least-squares (PLS) regression [6]. It is not exactly a validation but rather a calibration that allows to estimate a nitrogen content by referring to a database obtained by applying the Kjeldahl reference method to a set of samples close to the one being measured. Using this approach, the eventual bias of the Kjeldahl method is not detected and hence not corrected for.

## References


1 Borcic, V., Calmy, A., Hurst, S. et al. (2020). Buyers' club: une alternative pour l'accès au traitement? *Revue Médicale Suisse* 16: 2228–2231.

2 Bosch Reig, F. and Campíns Falcó, P. (1988). H-point standard additions method. Part 1. Fundamentals and application to analytical spectroscopy. *Analyst* 113: 1011–1016.

3 Wieczorek, M., Rengevicova, S., Świt, P. et al. (2017). New approach to H-point standard addition method for detection and elimination of unspecific interferences in samples with unknown matrix. *Talanta* 170: 165–172.

4 Guichard, N., Tobolkina, E., El Morabit, L. et al. (2021). Determination of antiretroviral drugs for buyers' club in Switzerland using capillary electrophoresis methods. *Electrophoresis* 42: 708–718.

5 Miller, J.N., Miller, J.C., and Miller, R.D. (2018). *Statistics and Chemometrics for Analytical Chemistry*, 6e. England: Pearson Education Limited.

6 Abdi, H., Chin, W.W., Vinzi, V.E. et al. (2013). *New Perspectives in Partial Least Squares and Related Methods*. Springer-Verlag.

# 11

# Conclusions

The role and interest of the measurements carried out and the results obtained, year after year, in the analytical laboratories are not denied. Both continuous improvement of analytical and bioanalytical methods and the importance of regulatory authorities to protect the population support this dynamic. Although it is difficult to put an exact figure on the number of results produced daily by the various laboratories, it is clear that it has been rapidly growing for several decades.

In the first chapter of this book, we have tried to explain the reasons for this growth, which are sometimes associated with a strong political and societal demand, but the response to this demand must be consistent.

The recent public health crisis associated with the SARS-CoV-2 virus (COVID-19) illustrates a fundamental requirement for the analytical sciences, namely the use of validated protocols and methods. For the first time in history, the administrative bodies in charge of public health have massively called upon laboratories to perform daily measurement to manage the pandemic. The success of this new strategy is open to question and should be evaluated.

The urgency of official demand has led analysts to a rush to use new methods with the risk of neglecting what had made their reputation since the 1990s, namely the standardization of procedures and inter-comparison of results and laboratories which are efficient in reducing measurement dispersion and discrepancies. During the first year of the crisis, numerous methods were developed and used without concern for harmonization. Whether to detect the virus and its variants, for example, by using reverse transcriptase polymerase chain reaction (RT-PCR) with several dozen probes that were proposed by various companies, or to quantify and/or detect the multiple antibodies produced by the virus, with the marketing of several dozen techniques and technologies in kit form.

The wealth of data produced has also raised many questions. For example, because of the lack of standardization, it was difficult after one year to use the first measurements obtained to understand the initial dynamics of the pandemic. As a result, by the end of 2020, several groups were already highlighting the risks resulting from the high frequency of false positives, but also of false negatives. Some even denounced part of the discourse about reinfections, which they explained precisely by measurement errors [1].

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

This example illustrates what we have been trying to explain throughout this book: an analytical method must always undergo serious, documented, and careful validation before being used routinely, and even more so when it is used to compare measurements made in different laboratories.

In addition, this validation must allow the analyst to explain clearly to the decision-maker the risk of considering the result as a definitive element of verdict. The concept of measurement uncertainty allows this risk to be identified and, more importantly, quantified. But, for measurement uncertainty to be able to play an effective role in decision-making based on an analytical result, a certain number of questions need to be asked. But a quick review will show that many answers are not too difficult to get.

## 11.1 Role of the Number of Replicates

It was pointed out in Section 8.4.3 that replication can reduce MU and facilitate or enable decision-making when the coverage interval appears overly large for an unambiguous decision. Making replicates is an effective approach to reduce uncertainty and thus facilitate decision-making. But it is necessary to be rigorous in how replication is designed to avoid any computational bias in estimating MU.

The role of replicates, in terms of number and mode of obtention, remains an open question. This is due to the ambiguity of the definition of a replicate in terms of independence from another replicate, but not only. Depending on the starting point of the replication within the sequence of the analytical operating procedure, as chosen by the analyst, the estimation of the MU is different.

When replication is conducted in the same series or done under repeatability conditions, the calculation is relatively simple. But it can also be done in another series (e.g. another day), and therefore acquired under intermediate precision conditions. From a statistical point of view it is mandatory to consider the between-series effect, which is often poorly known or estimated.

Another way to reduce an overly-large MU is to calculate the uncertainty budget. This tool is briefly discussed in Section 6.6. It consists in quantifying the relative importance of each source of uncertainty. This is straight-forward when applied to a type B approach. However for a type A approach, it implies a significant experimental effort and a complex statistical treatment, also mentioned in Section 7.2.4. Once the uncertainty budget is available, the main source(s) of uncertainty can be identified, and ways to reduce them can be explored.

## 11.2 Traceability to International Units

In Section 6.3, it was underlined that the attachment of chemical measurements to the International System of Units (SI) is delicate. Indeed, the mole never had a figurative standard, as exists for the meter or the kilogram. There has never been a

simple method to ensure its traceability. To overcome this peculiarity of the analytical sciences, the Consultative Committee for the Quantity of Matter (CCQM) has proposed two solutions: primary methods and certified reference materials.

To achieve traceability to SI units, primary analytical methods were recommended; as an example, an approach based on isotopic dilution coupled with high-resolution mass spectrometry (HR-MS). But there are also titrimetry and gravimetry following very particular protocols, which are admitted to the exclusive group of primary methods. For example analyzing pesticides by liquid chromatography mass spectrometry (LC-MS) or liquid chromatography hyphenated to tandem mass spectrometry (LC-MS/MS) using labeled molecules can lead to a primary method.

Certified reference materials are the other less expensive alternative, as they do not always require very sophisticated instrumentation; they are also promoted by the CCQM. Various protocols for interlaboratory comparisons are proposed. The following table (Table 11.1) is an attempt to classify these studies. The production of CRMs is considered the most satisfactory from a metrological point of view.

It can be said that a CRM is a substitute for a primary standard comparable to what the kilogram or the meter were in the past. Given the huge diversity of measurements in the analytical sciences, this helps to understand why unambiguous traceability will remain difficult.

**Table 11.1** Tentative classification of three interlaboratory comparison procedures.

| Characteristics | Collaborative study | Interlaboratory analysis | Proficiency testing |
|---|---|---|---|
| Main goal | Estimate the most probable value | Estimate method precision | Estimate laboratory competence |
| Context | Certified material production | Method standard publication | Laboratory accreditation |
| Main references | BIPM Guide, ISO 17034, FD ISO Guide 35 | ISO 5725 series | ISO 13528<br>ISO 17043 |
| Organizer | Official metrology laboratory | ISO expert groups, etc. | Certified organization |
| Number of participants | Less than 10 | At least 10 | Up to several thousand |
| Number of samples | 4 or 5 | 4 or 5 | Several/year |
| Number of measures/samples | Often 5 or 6 | At least 2 | 1 |
| Measurement method(s) | Combine as many methods as possible | Only one | Each lab has its own |

## 11.3 Education about Uncertainty

The practical problems that physicians, epidemiologists, researchers, analysts, engineers, and technicians solve every day have not a single solution but, in most cases, multiple solutions. The latter is documented by incomplete and uncertain data.

During high-school or university curriculum, to find the *unique* solution of an exercise, all useful elements are present in its statement. Unfortunately, this is not case in real life, and decisions must be made when only partial information is available. There is, therefore, a lack of training in measurement uncertainty in scientific curricula.

The classical (old) calculation of the maximum error, which used to be required of students after the sessions of practical work, can no longer be used anymore as a basis, considering the current evolution of knowledge in analytical science. Some initiatives exist to introduce this concept of MU at various levels of training, but it remains very fragmented. It can also be pointed out that the word "uncertainty" is inappropriate because it is confused with indecision, which is further emphasized by the famous precautionary principle.

However, we have tried to show how this parameter, inherent to any measurement, can reduce indecision. Will analysts soon report their results in the form, not as a single value but as a range of possible values, of a coverage interval? Throughout this book, we have presented this proposal as a necessary improvement for better interpretation of laboratory results and for decision-making. Paradoxically, end-users and prescribers may be reluctant and perceive this as inconvenient. It will therefore be necessary to explain and demonstrate the benefits to them and how they can integrate MU into their risk estimation.

## 11.4 Risk Analysis

Several times, we have argued that MU knowledge makes it easier to make decisions because it gives an idea of the risk involved. The practicality of this proposal was based on providing a result in the form of a coverage interval, defined as "containing the set of true values of a measurand with a given probability, based on the available information."

In all the examples presented, the coverage intervals included 95% of the possible true values of the measurand. In other words, this choice assumes that a percentage of possibly acceptable and *good* values are not considered as conforming. If the interval is used unilaterally, at most 2.5% of the values are rejected. Finally, the risk of a wrong decision would then be 2.5%. The question is to decide whether this risk is acceptable.

The choice of the 5% risk probability is very traditional. For example, a sampling plan is traditionally declared efficient when there is only a 5% chance of accepting a non-conforming lot. It is, therefore, a well-established industrial practice. But the use of analytical results is not limited to industrial batch release.

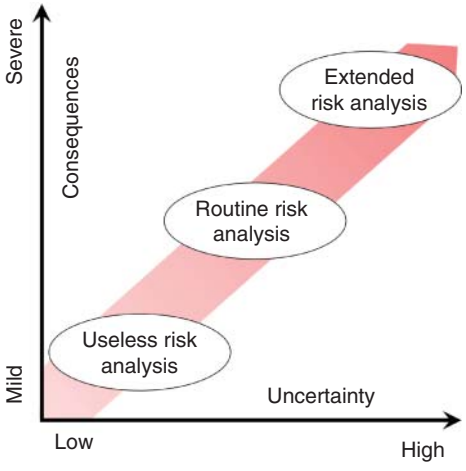



**Figure 11.1** Relationship between risk consequences and measurement uncertainty. Source: Adapted from Yoe [3].

For example, for official food control, sanitary limits are established with large safety margins, and this probability could be increased to 10% (bilaterally) without significant risk. But when it comes to a trial or a suspicion of doping and when the life or the reputation of a suspect is at stake, can we accept the risk of being wrong five times out of 100?

The starting point for considering consequences is risk analysis. It applies to many areas, such as health safety, natural disasters, public health, or occupational accidents, all areas where analytical sciences are involved. For example in the case of food risk analysis, there are standardized or official documents outlining the procedure to be followed [2]. In other fields, clear answers to this question must also be the subject of a consensus specific to the analytical field.

A possible guide for this reflection can be drawn from Figure 11.1 adapted from a book on decision-making under uncertainty [3]. As this reference is primarily concerned with economic decision-making, this figure is concerned with uncertainty in general, not just measurement uncertainty.

According to this scheme, if we take the example of the release of a batch based on an analytical measurement, it is up to the decision-maker to know, according to the MU provided by the laboratory – undoubtedly including the sampling uncertainty – what type of risk analysis he must conduct. It is then a matter of identifying the hazards and their effects and quantifying them in terms of probability.

## 11.5 Harmonization of MU Estimation Procedures

The last question that only analysts can answer is the reduction in the number of approaches proposed in the literature for estimating MU in the analytical sciences. It seems obvious that this plethora can introduce unfair commercial practices and confuse users. Throughout the numerous examples presented, we have tended to always relate to the method accuracy profile (MAP), developed in the 2000s. This

approach is empirical, because the organization of data collection and the data processing were initially intended for the method validation and finally turns out to be extremely informative.

It allows us to answer a complete set of questions such as how to estimate the analytical part of MU, which together with the sampling part, is often the major part of the MU. It also makes it possible to predict the number of measurements needed and so be able to budget for validation. Furthermore, we have also explained how, historically, operation procedure standardization has led to a profound improvement in confidence in analytical results. It seems obvious then that a similar approach can also be applied to the calculation of MU.

Several documents, standards, guides, and directives already exist, as listed in the bibliography. They should not be confused with the standards produced by the metrology organizations alone, which often concern physical measurements. However, unlike analytical method standards, which are always accompanied by an experimental contribution in the form of an interlaboratory comparison, standards describing MU estimation methods are not accompanied by a practical, collaborative, and comparative application.

In conclusion, improving the interpretation of analytical results still requires a lot of scientific and technical work. This book aims to make progress along this path, but it certainly does not claim to have reached the end. Only collaborative work, in which analysts know how to adopt these new concepts, has a chance of success. Moreover, many papers proposing the use of MU in decision-making remain theoretical and do not show the practical consequences of this approach. We have therefore tried to provide examples of how MU can be used in practice.

## References

1 Munoz Mendoza, J. and Alcaide, M.L. (2020). COVID-19 in a patient with end-stage renal disease on chronic in-center hemodialysis after evidence of SARS-CoV-2 IgG antibodies. Reinfection or inaccuracy of antibody testing. *IDCases* 22: e00943.

2 FAO and WHO (2019). *Codex Alimentarius Commission – Procedural Manual*, 27e, 254. Rome: http://www.fao.org/3/ca2329en/CA2329EN.pdf accessed 5 September 2023.

3 Yoe, C.E. (2019). *Principles of Risk Analysis: Decision Making under Uncertainty*. Boca Raton, FL: Taylor and Francis/CRC Press.

# Annexes

## The 10-step MAP Procedure

Step 1. Prepare a fully written Standard Operating Procedure (SOP) and define the analyte exactly, i.e. the quantity being measured.
Step 2. Select the bounds of the validation range and define the acceptance interval, which can be variable with the concentration.
Step 3. Select at least three *validation materials* representative of the validation range and the analyzed matrix. They are named *levels* and their exact contents must be known (they are the reference values) with an eventual uncertainty. They can be certified materials, surrogates, natural, or spiked matrices.
Step 4. Define the validation experimental design in terms of the number of *series* and replicates/per series. This is a crucial step that may condition the success of the study (see Section 5.4.2).
Step 5. If the method requires an external calibration, define the calibration experimental design.
Step 6. Run assays, collect measurements according to planned designs, and make a note of any discrepancies.
Step 7. If the method requires external calibration, estimate calibration curves for each series and calculate inverse-predicted concentrations.
Step 8. For each level (or validation material), calculate validation parameters, i.e. intermediate precision standard deviation and bias, with inverse-predicted concentrations.
Step 9. For each level, calculate the statistical dispersion interval containing the expected proportion of measurements as explained in Section 5.3.
Step 10. Interpret the results and define the validated range of the method, if any.

## Glossary of Used Terms

The vocabulary relating to the performance or validation of analytical methods is very abundant often confusing. For the purposes of this book, many definitions are derived from the International Vocabulary of Metrology published by the BIPM [1], where they exist. Some of the relevant definitions come from other documents and are given succinctly. Supplementary information can be found in the normative documents cited at the end of each chapter.

*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*,
First Edition. Max Feinberg and Serge Rudaz.

| Term | Definition |
|---|---|
| Acceptance interval | Interval of permissible measured values. Unless otherwise stated in the specification, the acceptance limits belong to the acceptance interval |
| Acceptance limit | Specified upper or lower bound of permissible measured values |
| Accuracy | Closeness of agreement between a measured value and a true value of a measurand |
| Accuracy profile | Combination, in graphical form, of one or more statistical dispersion intervals calculated at different concentration levels and one or more acceptance intervals |
| Analyte | Term used in analytical sciences instead of measurand |
| Between-series variance | Refers to the influence on the outcome of the organization of measures in the form of identifiable groups (related to ANOVA) |
| Bias | Estimate of a systematic error |
| Bound | Limit of an interval |
| Calibrant Surrogate Standard | Chemical compound used as a substitute for the authentic analyte for preparing calibration standards or calibrators |
| Calibration | Operation that, under specified conditions, in the first step, establishes a relation between the values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in the second step, uses this information to establish a relation for obtaining a measurement result from an indication |
| Calibration curve | Expression of the relation between indication and corresponding measured value |
| Calibrator Calibration Standard | Realization of the definition of a given quantity, with stated value and associated measurement uncertainty, used as a reference |
| Certified Reference Material (CRM) | Reference material, accompanied by documentation issued by an authoritative body and providing one or more specified property values with associated uncertainty and traceability, using valid procedures |
| Characteristic | An entity or concept that is a distinctive or specific feature of a method, such as robustness or accuracy |
| Combined standard uncertainty | Standard measurement uncertainty that is obtained using the individual standard measurement uncertainties associated with the input quantities in a measurement model |
| Conformance probability | Probability that an item fulfills a specified requirement |
| Conformity assessment | Activity to determine whether specified requirements relating to a product, process, system, person, or body are fulfilled |

| Term | Definition |
|---|---|
| Consumer's risk | Probability that a particular accepted item is non-conforming |
| Conventional value | Value attributed by agreement to a quantity for a given purpose |
| Correction | Compensation for an estimated systematic effect |
| Coverage factor | Number larger than one by which a combined standard measurement uncertainty is multiplied to obtain an expanded measurement uncertainty |
| Coverage interval | Interval containing the set of true values of a measurand with a stated probability, based on the information available |
| Coverage probability | Probability that the set of true values of a measurand is contained within a specified coverage interval |
| Decision rule | Documented rule that describes how measurement uncertainty will be accounted for about accepting or rejecting an item, given a specified requirement and the result of a measurement |
| Definitional uncertainty | Component of measurement uncertainty resulting from the finite amount of detail in the definition of a measurand |
| Effect | ANOVA: contribution of a specific factor level to the output variable |
| Error | Measured value minus a reference value |
| Estimator | In statistics, it is a rule that creates an estimate from observed data. For example, the sample mean is an estimator for the theoretical mean |
| Evaluation/Characterization | Study of the qualities of a process, technique, or instrument to specify its characteristics and its adaptation to the intended purpose |
| Expanded uncertainty | Product of a combined standard measurement uncertainty and a factor larger than the number one |
| Fitness-for-purpose | Concept covering the level of appreciation of the ability of the data obtained by a measurement process to enable a user to make technically and administratively sound decisions |
| Guard band | Interval between a tolerance limit and a corresponding acceptance limit |
| Input quantity | Quantity that must be measured, or a quantity, the value of which can be otherwise obtained, to calculate a measured value of a measurand |
| Inspection | Conformity assessment by observation and judgment accompanied, as appropriate, by measurement, testing, or gauging |
| Interlaboratory study | Organization, execution, and evaluation of measurements or tests on the same or similar entities by two or more laboratories under predetermined conditions |

| Term | Definition |
|---|---|
| Intermediate precision condition | Condition of measurement, out of a set of conditions that includes the same measurement procedure, same location, and replicate measurements on the same or similar objects over an extended period, but may include other conditions involving changes |
| Intermediate precision variance | Measurement precision under a set of intermediate precision conditions of measurement |
| International System of Units | System of units, based on the international system of quantities, their names, and symbols, including a series of prefixes and their names and symbols, together with rules for their use, adopted by the General Conference on Weights and Measures (CGPM) |
| Interval | Interval is used together with the symbol $[a, b]$ to denote the set of real numbers $x$ for which $a \leq x \leq b$, where $a$ and $b$ are real numbers. The term is used here for closed intervals. The symbols $a$ and $b$ denote the endpoints of the interval $[a, b]$ |
| Limit of detection | Measured value, obtained by a given measurement procedure, for which the probability of falsely claiming the absence of a component in a material is, given a probability of falsely claiming its presence |
| Limit of quantification (1) | Smaller and/or greater concentration of the analyte that can be quantified, under the experimental conditions described in the method |
| Limit of quantification (2) | Smaller and/or larger quantity of the analyte in a sample that can be dosed under the experimental conditions described with defined uncertainty. It corresponds to the smallest and/or largest concentration of the validated range |
| Limit of quantification (3) | The lowest concentration of an analyte that can be quantified in a sample with acceptable risk of error, under indicated operating conditions |
| Limit of quantification (4) | The lowest concentration for which a coefficient of variation of repeatability equal to a given threshold is obtained |
| Linearity | Establishment that there is a linear relationship between the quantities found (or quantified) in samples and their reference values |
| Matrix | A set of constituents forming the sample, other than the researched analyte. Matrix effects (or matrix interferences) reflect the possible influence the matrix constituents may have on the measuring device response |
| Maximum permissible error | For a measuring instrument, maximum difference, permitted by specifications or regulations, between the instrument indication and the quantity being measured |

| Term | Definition |
|---|---|
| Measurand | Quantity intended to be measured (see Analyte) |
| Measured value | Value representing a measurement result |
| Measurement | Process of experimentally obtaining one or more values that can reasonably be attributed to a quantity |
| Measurement capability index | Tolerance divided by a multiple of the standard measurement uncertainty associated with the measured value of a property of an item. Generally taken to be four |
| Measurement procedure | Detailed description of a measurement according to one or more measurement principles and to a given measurement method, based on a measurement model, and including any calculation to obtain a measurement result |
| Measurement uncertainty (MU) | Non-negative parameter characterizing the dispersion of the values being attributed to a measurand based on the information used |
| Measurement method | Generic description of a logical organization of operations used in a measurement |
| Measurement model | Mathematical relation among all quantities known to be involved in a measurement |
| Measurement result | Set of values being attributed to a measurand together with any other available relevant information |
| Measuring interval or working interval | Set of values of quantities of the same kind that can be measured by a given measuring instrument or measuring system with specified instrumental uncertainty, under defined conditions |
| Measuring transducer | Device, used in measurement, that provides an output quantity having a specified relation to the input quantity |
| Nominal property | Property of a phenomenon, body, or substance, where the property has no magnitude |
| Output quantity | The measured value of which is calculated using the values of input quantities in a measurement model |
| Parameter | Statistical model coefficients or combination of coefficients are used to convey the performance of a method. They are random variables |
| Precision | Closeness of agreement between indications or measured values obtained by replicating measurements on the same or similar objects under specified conditions |
| Primary measurement procedure | Reference measurement procedure used to obtain a measurement result without relation to a measurement standard for a quantity of the same kind |
| Procedure | Operations to be carried out, precautions to be taken, and measures to be applied are contained in documents specific to each laboratory |
| Producer's risk | Probability that a conforming item will be rejected based on a future measurement result |

| Term | Definition |
|---|---|
| Qualification | Operation to demonstrate that an analytical system or instrument is functioning properly and yielding the expected results |
| Quality Control | Sample adapted to the method used and intended to assess the accuracy and precision of the results |
| Quantity | Property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference |
| Random error | Component of measurement error that, in replicate measurements, varies in an unpredictable manner |
| Range | Range of the interval $[a, b]$ is the difference $b - a$, and is denoted by $r[a, b]$ |
| Recovery yield | Percentage of the known or assigned concentration of an analyte recovered thanks to the analytical procedure |
| Reference (or official) measurement procedure | Measurement procedure accepted as providing measurement results fit for their intended use in assessing measurement trueness |
| Reference material (RM) | Material, sufficiently homogeneous and stable with reference to specified properties, that has been established to be fit for its intended use in measurement or in examination of nominal properties |
| Reference value | Value used as a basis for comparison for values of the same nature |
| Rejection interval | Interval of non-permissible measured values |
| Relative standard uncertainty | Standard measurement uncertainty divided by the absolute value of the measured value |
| Repeatability | Measurement precision under a set of repeatability conditions of measurement |
| Repeatability condition | Condition of measurement, out of a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions, and same location, and replicate measurements on the same or similar objects over a brief period |
| Reproducibility | Measurement precision under a set of reproducibility conditions of measurement |
| Reproducibility condition | Out of a set of conditions that includes various locations, operators, measuring systems, and replicate measurements on the same or similar objects |
| Sampling | Act permitting the obtaining of a biological sample or entity taken or, by extension, the result of that act |
| Sensitivity | Quotient of the change in an indication of a measuring system and the corresponding change in a value being measured |
| Series | Set of measurements, replicated or not, performed under repeatability condition |

| Term | Definition |
|---|---|
| Specification Specified requirement | Need or expectation that is stated. Specified requirements may be stated in normative documents such as regulations, standards, and technical specifications. A typical specified requirement takes the form of a stated interval of permissible values of a measurable property of an item |
| Specification limit | Specified upper or lower bound of permissible values of a property |
| Specificity | Capability of a measuring system or operating procedure to measure the concentration of a given analyte |
| Standard addition or Spike | Addition to a material or standard of chemical composition defined with uncertainty in a known quantity. Conventionally, this operation is intended to confirm the trueness of a method or to calibrate an instrument |
| Standard uncertainty | Measurement uncertainty expressed as a standard deviation |
| Statistical dispersion interval | Interval determined from a random sample so that there is a specified level of confidence that the interval covers at least a given proportion of the sampled population |
| Systematic error | Component of measurement error that in replicate measurements remains constant or varies in a predictable manner |
| Target uncertainty | Measurement uncertainty specified as an upper limit and decided based on the intended use of measurement results |
| Tolerance interval (conformity assessment) | Interval of permissible values of a property. Unless otherwise stated in a specification, the tolerance limits belong to the tolerance interval. This term used in conformity assessment has a different meaning than as used in statistics |
| True value | Value consistent with the definition of a quantity |
| Trueness | Closeness of agreement between the average of an infinite number of replicate measures of a value and a reference value |
| Type A evaluation | Evaluation of a component of measurement uncertainty by a statistical analysis of measured values obtained under defined measurement conditions |
| Type B evaluation | Evaluation of a component of measurement uncertainty determined by means other than type A evaluation |
| Uncertainty budget | Statement of a measurement uncertainty, of the components of that measurement uncertainty, and of their calculation and combination |
| Validation | Verification, where the specified requirements are adequate for an intended use |
| Validation criterion | Element to which one refers to judge, assess, or define whether a method is valid |

| Term | Definition |
|---|---|
| Validation range | Concentration range and matrix types covered by the validation study |
| Validation standard | Material selected as representative of the matrix to which the validation study relates. It must be homogenized and of known concentration. In the most favorable case, it is a RM sometimes certified CRM |
| Verification | Provision of objective evidence that a given item fulfills specified requirements |
| Within-series variance | Refers to the influence of replicate measures on identifiable groups (related to ANOVA) |
| Working interval | Set of values of quantities of the same kind that can be measured by a given measuring instrument or measuring system with specified instrumental measurement uncertainty, under defined conditions |
| β-Expectation Tolerance interval | Statistical dispersion interval, with an expected coverage proportion $\beta\%$ of possible values |
| β-γ-Content Tolerance interval | Statistical dispersion interval, with a coverage proportion $\beta\%$ of values, with a confidence level $\gamma\%$ |

## Acronyms

| Acronym | Expanded name |
|---|---|
| A2LA | American Association for Laboratory Accreditation |
| AFNOR | Association Française de Normalisation |
| AIDS | Acquired immunodeficiency syndrome |
| ANOVA | Analysis of variance |
| AOAC | Association of Official Analytical Chemists |
| API | Active principal ingredient |
| BIPM | Bureau International des Poids et Mesures |
| CAC | Commission of the Codex Alimentarius |
| CCMAS | Codex Committee on Methods of Analysis and Sampling |
| CCQM | Consultative Committee on the Quantity of Matter |
| CI | Confidence interval |
| CIPM | Convention Internationale des Poids et Mesures |
| COFRAC | Comité Français d'Accréditation |
| COMAR | Code of reference materials |
| CRM | Certified reference material |
| *df* | Degrees of freedom |
| EC | External calibration |

| Acronym | Expanded name |
|---|---|
| EDQM | European Directorate for the Quality of Medicines |
| EMA | European Medicines Agency |
| EPA | U.S. Environmental Protection Agency |
| ERM | External reference material |
| FDA | Food and Drug Administration |
| GEON | OMCL (official medicines control laboratory) network of the Council of Europe |
| GUM | Guide to the expression of uncertainty in measurement |
| HPSAM | H-point standard addition method |
| HR-MS | High-resolution mass spectrometry |
| IC | Internal calibration |
| ICH | Before 2015: International Conference on Harmonization of technical requirements for registration of pharmaceuticals for human use |
| ICH | International Council for Harmonization of technical requirements for registration of pharmaceuticals for human use |
| ICP-ID-MS | Isotope dilution inductively coupled plasma mass spectrometry |
| IEC | International Electrotechnical Commission |
| IFCC | International Federation of Clinical Chemistry and Laboratory Medicine |
| ILAC | International Laboratory Accreditation Cooperation |
| IPD | Isotopic pattern deconvolution |
| IRM | Internal reference material |
| IS | Internal standard |
| ISC | In-sample calibration |
| ISO | International Organization for Standardization |
| IUPAC | International Union of Pure and Applied Chemistry |
| IUPAP | International Union of Pure and Applied Physics |
| JCGM | Joint Committee for Guides on Metrology |
| LC | Liquid chromatography |
| LC-MS | Liquid chromatography–mass spectrometry |
| LC-MS/MS | Liquid chromatography coupled to tandem mass spectrometry |
| $LD_{50}$ | Level of detection at 50% |
| LIMS | Laboratory information management system |
| LLOQ | Lower limit of quantification |
| LOD | Limit of detection |
| LOQ | Limit of quantification |
| MAP | Method accuracy profile |
| MIRM | Multiple isotopologue reaction monitoring |
| MRL | Maximum residue limit |

| Acronym | Expanded name |
|---|---|
| MS | Mass spectrometry |
| MU | Measurement uncertainty |
| NIRS | Near-infrared spectrometry |
| OIML | Organisation Internationale de Métrologie Légale |
| OLS | Ordinary least-squares |
| OMCL | Official medicines control laboratory |
| OOS | Out of specification |
| QC | Quality control |
| RF | Response factor |
| RM | Reference material |
| *RSD* | Relative standard deviation or coefficient of variation |
| SAM | Standard addition method |
| SEC | Exclusion chromatography or molecular sieve chromatography |
| SFSTP | Société Française des Sciences et Techniques Pharmaceutiques |
| SI | International System (of units) |
| SIL | Stable isotope labeled |
| SMPR | Standard method performance requirement |
| SPC | Statistical process control |
| std. dev. | Standard deviation |
| TDI | Total daily intake |
| TI | Tolerance interval |
| TRV | Toxicological reference value |
| *ULOQ* | Upper limit of quantification |
| USP | US pharmacopeia |
| UV | Ultraviolet |
| VIM | International Vocabulary of Metrology |
| WADA | World Anti-doping Agency |
| WHO | World Health Organization |
| WLS | Weighted least-squares |
| β-ETI | β-Expectation tolerance interval |
| β-γ-CTI | β-γ-Content tolerance interval |

## Reference

1 BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML (2012). *International Vocabulary of Metrology — Basic and General Concepts and Associated Terms (VIM3)*. Sèvres, France: JCGM 200, https://www.bipm.org/ (accessed 23 July 2023).

# Index


## *a*

absolute inverse-predicted concentrations 120, 121
absorbance 6, 18, 25, 161, 240, 278, 281
active principal ingredient (API) 271
Akaike information coefficient (AIC) 132
analysis of variance (ANOVA) 13, 65–72, 98, 187
analytical validation 55, 227
ANOVA classic algorithm 72–75
a posteriori 136, 192
authentic analyte in surrogate matrix 12–13
Avogadro's constant 152

## *b*

balanced and unbalanced experimental design 71–72, 125
Beer's law 6, 25
β-expectation tolerance interval (β-ETI) 114, 118, 119, 124–128, 237, 257, 263, 302
β-γ content tolerance interval (β-γ-CTI) 125, 128
between-laboratories variance 61, 69
between-series variance 61, 67, 69, 72, 136–138, 279, 281, 296
bias and recovery yield 82–83
Bioanalytical Method Validation 142

## *c*

calibration
  direct and inverse 25–28
  least-squares regression method 28–34
  metrological approach 51–53
  misuses of regression 45–48
  modes 9–10
  nonlinear calibration curve 41–44
  ordinary least-squares (OLS) regression 34–37
  quantification and 2–3
  standard addition method (SAM) 48–51
  weighted least-squares (WLS) regression 37–40
calibrators 3–5, 12, 14, 16, 26–30, 32, 33, 36–38, 43, 52–56, 153, 182, 265
capillary zone electrophoresis (CZE) 273
certified reference material (CRM) 7, 87, 153, 156, 188, 191, 193, 291, 296
check material homogeneity/stability 91, 97–99
chromatographic assays (CC) 142
classic validation procedure 109, 110
coefficient of correlation 45, 46
coefficient of determination 35, 36, 40, 45–47, 52, 132
composed standard uncertainty 151


*Quantification, Validation and Uncertainty in Analytical Sciences: An Analyst's Companion*, First Edition. Max Feinberg and Serge Rudaz.

conformity assessment 89, 124, 211, 216, 217, 227–231, 250, 263, 296, 297, 301
constitutional heterogeneity coefficient 234
control chart 87, 99–102, 114, 125, 128, 144, 147, 160, 186, 191–200
coverage interval 151
  of given concentration 215
  of given relative uncertainty 215–216
  of limits 216–217
  oforigin 211–215

## d

decision limit and detection capability
  definitions 262–264
  example of calculation 267–269
  initial procedure 265
  modified procedure 265–267
decision rule, defined 227–229
decision *versus* uncertainty 221–223
deuterium effect 8
direct calibration 27, 28
direct internal calibration 20–23
distributional heterogeneity 234
diverse precision parameters 59–60
Dumas method 284–287

## e

education about uncertainty 292
error of measurement 149
estimators by statisticians 108
expanded uncertainty 89, 151, 169–171, 189–191, 206, 229–231, 282, 297
expected relative uncertainty 260–262
external calibration (EC) 9–14, 23, 25, 28, 49, 54, 55, 271, 295
external reference materials (ERM) 87, 188

## f

figures-of-merit 108
fitness-for-purpose 223, 262, 297
framework for decision-making
  decision *versus* uncertainty 221–223
  specification limits and reference values 223–226
fundamental uncertainty 222

## g

generic calibration curve 27
generic cause to effect diagram 182–184
generic measurement model 181–182, 184, 187, 189, 201
gold standard method 191, 283
guard band 225, 229–231, 266, 297

## h

harmonization of MU estimation procedures 293–294
homogeneity check, procedure 236–237
H-point standard addition method (HPSAM) 17, 272, 283

## i

inconsistencies of validation vocabulary 107–109
in-sample calibration (ISC) 10, 11, 15–17, 26, 272
interlaboratory comparison data 200–203
intermediate precision condition 60, 61, 65, 70–72, 113, 135, 159, 160, 181, 182, 187, 229, 237, 251, 253–254, 264, 275, 290, 298
intermediate precision variance 61, 69, 127, 182, 188, 298
internal calibration 10, 18, 21, 23, 26
internal reference material (IRM) 8, 87, 88, 160, 188, 193, 194
intra-laboratory standard deviation 160
inverse calibration 25–29, 33, 35, 36, 258
isotope dilution mass spectrometry (IDMS) 18–20
isotope inversion 14
isotopic pattern deconvolution (IPD) 11, 18–20, 36

## k

Kjeldahl and Dumas method comparison 285–287
Kragten iterative algorithm 165–169

## l

law of propagation of uncertainty 161–164, 181, 184, 187, 199, 201, 244, 253
least-squares regression method 28–34, 110
ligand binding assays (LBAs) 142
limit of detection (LOD) 106, 170, 257, 298
limit of quantification (LOQ)
  assessment of 258
  definitions of 258
  expected relative uncertainty 260–262
linearity, defined 47–48

## m

matrix-matched external calibration (MMEC) 10, 14, 55
maximum residue limits (MRL) 206, 225, 263, 264
measurand (Analyte) 1–9, 154–157
measurement uncertainty (MU) 1, 30, 272
  accuracy, total error, and uncertainty 171–173
  expanded uncertainty 169–170
  general procedure 150–151
  insight on probability 174–177
  Kragten iterative algorithm 165–169
  law of propagation of uncertainty 161–164
  measurand 154–157
  principle of 149–150
  rounding of results 170–171
  traceability at the International System of units 152–154
  uncertainty components 157–158
  uncertainty sources 158–161
measurement uncertainty and decision
  calibration model 240–243
  framework for decision-making
    analytical report role 226–227
    decision *versus* uncertainty 221–223
    specification limits and reference values 223–226
  number of replicates 250–251
  replication under intermediate precision condition 253–254
  replication under repeatability condition 251–253
  sample conformity assessment
    decision rule 227–229
    guard band concept 229–231
  sampling uncertainty
    example of copper in wheat flour 237–239
    and heterogeneity 232–236
    homogeneity check procedure 236–237
  uncertainty of corrected results 243–249
measurement uncertainty in analytical sciences
  control charts data 191–200
  coverage interval 211–217
  interlaboratory comparison data 200–203
  method accuracy profile data 181–191
  published procedures 179–181
  uncertainty functions 203–211
Metre Convention 149
method accuracy profile (MAP) 7, 26, 56, 70, 89, 107, 112–122, 156, 181–191, 241, 243, 261, 281–283, 294
method comparison using uncertainty
  analyte defined by the operating procedure 283–284
  Kjeldahl and Dumas method comparison 285–287
method validation 227
  accuracy profile 131–145

method validation (*contd.*)
method accuracy profile (MAP) 113–122
review of validation procedures 105–113
statistical dispersion intervals 122–131
metrology 1, 2, 51, 81, 86–89, 108, 149, 150, 188, 212, 257, 294, 295
misuses of regression 45–48
multiple blank standard deviations 258–259
multiple isotopologue reaction monitoring (MIRM) 21
multipoint matrix-matched external calibration (MMEC) 55

***n***

nonlinear calibration curve 41–44, 47
normalized response ratio 8
*n*-way ANOVA 187

***o***

one-factor ANOVA 187
one-way fixed effects ANOVA 64
one-way random effects ANOVA 64, 236
optical density (OD) 25, 43, 161, 274
ordinary least-squares (OLS) 29, 30, 34–37, 67
organization of proficiency testing schemes 90–91

***p***

partial least-squares (PLS) 7, 53, 110, 287
performance scores 93–94, 108
precision
analysis of variance (ANOVA) 64–75
ANOVA classic algorithm 72–75
balanced and unbalanced experimental design 71–72
detect outliers and stragglers 75–78
diverse precision parameters 59–60
prediction interval 123, 124, 176, 213
proficiency testing 89, 160
proficiency testing scheme (PTS) 79, 89–91, 98, 200, 236
proportion of nonacceptable measures 145–147

***q***

quality control (QC) 46, 90, 99, 100, 102, 125, 142–144, 147, 160, 191, 193, 226, 227, 271, 273, 300
quantification
authentic *versus* surrogate 3–6
calibration 2–3, 9–10
direct internal calibration with labeled calibrant 20–23
external calibration (EC) 10–14
in-sample calibration (ISC) 15–17
isotopic pattern deconvolution (IPD) 18–20
signal pretreatment and normalization 6–9

***r***

random effect factor 73, 125, 187
reference materials 18, 56, 87–88, 109, 134, 153, 156, 188, 193, 194, 199, 291, 296, 300
reference value
of test material 91–93
value uncertainty 198–200
relative standard deviation (RSD) 20, 69–70, 82, 117, 173, 204, 205, 258
relative uncertainty 151, 153, 170, 173, 176, 190, 199, 208–211, 215–217, 224, 230, 231, 238, 239, 242, 245, 248, 249, 260–262, 277, 278, 282, 285
repeatability condition 59, 60, 64, 72, 98, 113, 135, 182, 236, 237, 251–253, 275–277, 290, 300
repeatability variance 61, 62, 69, 72, 136, 201, 202, 276, 279
reproducibility condition 60, 63, 68, 300
reproducibility variance 8, 9, 63, 69, 202
Resource H β-expectation tolerance interval (Excel) 117
risk analysis 222, 225, 292–293
robust estimators 78, 92, 95
role of replicates 290

**s**

sample conformity assessment 89, 216, 217, 227–231, 263
sampling uncertainty
  example of copper in wheat flour 237–239
  homogeneity check procedure 236–237
  sampling and heterogeneity 232–236
SARS-CoV-2 virus (COVID-19) 289
scientific uncertainty 222
sensitivity coefficient 164, 168, 182, 201
Shewhart control chart 101, 191–195, 197
signal-to-noise ratio 259, 262, 266
sources of uncertainty in the laboratory
  machine/equipment 186
  manpower 185
  material and handling of items 185
  measurement and other sources 186–187
  method 185–186
stable isotope-labeled (SIL)
  chemicals 8
  molecule 13
standard addition method (SAM) 15, 48, 272
  estimation from method accuracy profile 281–283
  without replication 273–279
  with replication 279–281
  and surrogate samples 88–89
standard deviation of grand mean 72
standard uncertainty 89, 92, 151, 152, 158, 159, 161–163, 165–167, 169, 170, 172, 173, 183, 184, 187–189, 191, 195, 199–202, 206, 208–211, 215, 238, 244, 245, 251–253, 269, 276, 296, 300, 301
statistical dispersion intervals 112–114, 122–131, 141, 143, 195–198, 212, 295, 296, 301, 302
statistical process control (SPC) 99, 123, 193
straight-line computation 28–31
surrogate calibrant in authentic matrix 13–14
surrogate standard 3–5, 13, 14, 20, 21, 56, 296

**t**

10-step MAP procedure 114, 295
theophylline 26–28, 36, 39, 40, 116–121, 127, 129–133, 136, 138–141, 143–146, 152, 153, 155, 190–192, 208–210, 215, 241–243, 252, 253, 261
total analytical error (TAE) 111, 112, 149, 150, 173, 190
toxicological reference values (TRV) 225
traceability
  International System of units 152–154
  to International Units 290–291
transmittance 25
trueness
  algorithm A 94–97
  bias and recovery yield 82–83
  check material homogeneity/stability 97–99
  control charts 99–102
  evolution of 83–84
  organization of proficiency testing schemes 90–91
  performance score 93–94
  primary operating procedures 86
  reference materials 87–89
  reference value of 91–93
  specificity and sources of bias 84–86

**u**

uncertainty components 56, 157–158, 160, 164
uncertainty functions 57, 113, 168, 176, 177, 182, 203–211, 215–217, 224, 229, 231, 236, 238, 242, 243, 245, 248, 249, 251–253, 260, 261, 269, 285, 286
uncertainty sources 53, 56, 158–161, 183
upper limit of quantification (ULOQ) 12, 120, 260

**V**
validation plans 109–113, 281
variance of grand mean 70, 189, 279
variance ratio 71, 126, 127, 136–140, 146, 237
verification standard solutions 8

**W**
weighted least-squares (WLS) 37–40, 110, 133
World Anti-Doping Agency (WADA) 20

# WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.