



中华人民共和国国家标准

GB/T 15000.3—2023/ISO Guide 35:2017

代替 GB/T 15000.3—2008

标准样品工作导则 第3部分：标准样品 定值和均匀性与稳定性评估

Directives for the work of reference materials—
Part 3: Reference materials—Characterization and
assessment of homogeneity and stability

(ISO Guide 35:2017, Reference materials—
Guidance for characterization and assessment of
homogeneity and stability, IDT)

2023-05-23 发布

2023-09-01 实施

国家市场监督管理总局 发布
国家标准化管理委员会

目 次

前言	III
引言	V
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 符号	3
5 约定	4
6 标准样品生产概述	4
6.1 概述	4
6.2 项目设计概要	4
6.3 原材料的获取	5
6.4 可行性研究	6
6.5 标准样品加工	6
6.6 均匀性评估	6
6.7 稳定性评估	6
6.8 测量程序的选择	6
6.9 计量溯源性	6
6.10 定值和不确定度评定	6
6.11 互换性评估	7
6.12 运输事宜	7
6.13 赋值	7
6.14 稳定性监测	7
6.15 重复批生产的标准样品	7
7 均匀性评估	7
7.1 概述	7
7.2 均匀性试验研究的需求	8
7.3 拟研究的特性	9
7.4 统计有效的抽样计划	9
7.5 均匀性研究的测量程序选择和实施	11
7.6 均匀性研究设计	12
7.7 评定均匀性研究	15
7.8 测量程序重复性欠佳的情况	17
7.9 单元内均匀性	18

7.10	均匀性的确认	20
7.11	均匀性研究的不确定度评定	20
8	稳定性评估与监测	21
8.1	概述	21
8.2	稳定性评估	22
8.3	稳定性研究的分类	24
8.4	有效稳定性研究的一般要求	25
8.5	评定稳定性研究结果	26
8.6	稳定性研究中发现显著趋势后的措施	32
8.7	稳定性研究的不确定度评定	32
8.8	由稳定性研究估计储存寿命(“保存期”)	34
8.9	有关稳定性管理的使用说明	34
8.10	稳定性监测	34
9	材料定值	37
9.1	概述	37
9.2	建立计量溯源性	37
9.3	在单个实验室中采用单一参考测量程序(参见 ISO/IEC Guide 99 定义)进行定值	39
9.4	在一个或多个有能力的实验室采用两种或两种以上可证明准确度的方法对非操作定义的被测量进行定值	41
9.5	采用有能力的实验室网络对操作定义的被测量进行定值	43
9.6	纯度	44
9.7	身份	45
9.8	存在/不存在	47
9.9	定序尺度	47
9.10	定性特性	47
9.11	非标准值的定值	47
10	测量不确定度评定	48
10.1	评定 CRM 特性值不确定度的基础	48
10.2	批定值的基本模型	48
10.3	不确定度来源	48
10.4	包含区间和因子	49
附录 A (资料性)	定值研究的设计与评价	50
附录 B (资料性)	统计方法	58
附录 C (资料性)	实例	66
附录 D (资料性)	测量不确定度评定	74
参考文献		75

前 言

本文件按照 GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件是 GB/T 15000《标准样品工作导则》的第3部分。GB/T 15000 已经发布了以下部分：

- 第1部分：在技术标准中陈述标准样品的一般规定；
- 第2部分：常用术语及定义；
- 第3部分：标准样品 定值和均匀性与稳定性评估；
- 第4部分：证书、标签和附带文件的内容；
- 第5部分：质量控制样品的内部研制；
- 第6部分：标准样品包装通则；
- 第7部分：标准样品生产者能力的通用要求；
- 第8部分：标准样品的使用。

本文件代替 GB/T 15000.3—2008《标准样品工作导则(3) 标准样品定值的一般原则和统计方法》，与 GB/T 15000.3—2008 相比，除结构调整和编辑性改动外，主要技术变化如下：

- 更改了标准的范围(见第1章,2008年版的第1章)；
- 增加了“测量模型”“标准值”“信息值”“校准样品”“质量控制样品”“同步稳定性研究”和“〈标准样品的〉生产”等7个术语(见3.3、3.5、3.6、3.7、3.8、3.9、3.10)；
- 删除了“测定”“瓶间均匀性”“瓶内均匀性”“混合”“基体材料”“加料”“短期稳定性”“长期稳定性”“使用寿命”和“有效期”等10个术语(见2008年版的3.4、3.5、3.6、3.7、3.8、3.9、3.10、3.11、3.12、3.13)；
- 更改了“标准样品”和“有证标准样品”2个定义(见3.1、3.2,2008年版的3.1、3.2)；
- 增加了“约定”“定值研究的设计与评价”“测量不确定度评定”等内容的有关规定(见第5章、附录A、附录D)；
- 将“项目设计”更改为“标准样品生产概述”，细化并更改了“计量溯源性”“定值和不确定度评定”“互换性评估”“赋值”“稳定性监测”“重复批生产的标准样品”等相应的技术内容；(见第6章,2008年版的第5章)；
- 将“均匀性研究”更改为“均匀性评估”，细化并更改了“统计有效的抽样计划”“均匀性研究的测量程序选择和实施”“均匀性的确认”“均匀性研究的不确定度评定”等相应的技术内容(见第7章,2008年版的第7章)；
- 将“稳定性研究”更改为“稳定性评估和监测”，细化并更改了“稳定性评估”“稳定性研究的分类”“有效稳定性研究的一般要求”“稳定性研究中发现显著趋势后的措施”“稳定性研究的不确定度评定”等相应的技术内容(见第8章,2008年版的第8章)；
- 将“特性值测定”更改为“材料定值”，“建立和证明溯源性”更改为“建立计量溯源性”，同时细化了实用方法、测量方案设计和有关特性的一些问题等技术内容(见第9章,2008年版的第9章)；
- 将“测量不确定度评估”更改为“测量不确定度评定”，细化并更改了“评定 CRM 特性值不确定度的基础”“包含区间和包含因子”等相应的技术内容(见第10章,2008年版的第6章)；
- 删除了“数据及不确定度评估”“证书”等内容的有关规定(2008年版的第10章、第11章)。

本文件等同采用 ISO Guide 35:2017《标准样品 定值和均匀性与稳定性评估指南》。文件类型由 ISO 的指南文件调整为我国的国家标准。

本文件做了下列最小限度的编辑性改动：

——为与现有标准协调，将标准名称改为《标准样品工作导则 第3部分：标准样品 定值和均匀性与稳定性评估》。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由全国标准样品技术委员会(SAC/TC 118)提出并归口。

本文件起草单位：中国标准化协会、冶金工业信息标准研究院、建筑材料工业技术监督研究中心、国家地质实验测试中心。

本文件主要起草人：徐大军、吴忠祥、卢行安、何平、王向红、田文、唐本玲、赵亚娴、张帆、石雨婷、胡晓燕、刘英、徐建平。

本文件及其所代替文件的历次版本发布情况为：

——1994年首次发布为GB/T 15000.3—1994；

——2008年第一次修订时，并入GB/T 15000.5—1994《标准样品工作导则(5) 化学成分标准样品技术通则》的内容(GB/T 15000.5—1994的历次发布版本情况为：GB/T 15000.5—1994)；

——本次为第二次修订。

引 言

标准样品(RM)的生产是改善和维持全球测量一致性体系的关键活动。正如 ISO 指南 33^[1]所述,具有不同特性的 RM 被应用于测量,如校准、质量控制、能力验证和方法确认,也用于给其他材料赋值。有证标准样品(CRM)还用于确认或建立诸如辛烷值、硬度和 pH 等约定标尺的计量溯源性。为了能在时空上可比,测量需要可追溯至适当的和规定的参考标准。在化学、生物学和物理学以及涉及物质和材料的其他学科中,CRM 在建立测量结果溯源性方面具有重要作用。实验室应用 CRM 作为易于获得的测量标准,以建立其测量结果与国际标准的可溯源性。CRM 所具有的特性值可以追溯到国际单位制(SI)或生产过程中其他国际公认的参考标准。在这方面,我国已经建立了支撑标准样品制作和应用的基础性国家标准体系,以确保其生产的标准样品质量符合最终使用者的要求。在该标准体系中,GB/T 15000《标准样品工作导则》是指导我国标准样品工作的基础性和通用性标准。GB/T 15000旨在确立普遍适用于标准样品生产者制备、定值、包装储存、分发和应用工作的准则,由 8 个部分构成。

- 第 1 部分:在技术标准中陈述标准样品的一般规定。目的在于规定技术标准制定中陈述标准样品的一般要求、标准样品与技术标准的关系、选择标准样品的技术准则等。
- 第 2 部分:常用术语及定义。目的在于规定与标准样品有关的术语和定义,特别是用于标准样品产品说明书、证书和相应的研制报告中的术语。
- 第 3 部分:标准样品 定值和均匀性与稳定性评估。目的在于就技术问题提供更具体的指导,并提供标准样品均匀性、稳定性评估和定值过程的基本模型。
- 第 4 部分:证书、标签和附带文件的内容。目的在于确立适用于编写标准样品证书、标签和附带文件内容时需要遵守的相关要求。
- 第 5 部分:质量控制样品的内部研制。目的在于规范和指导质量控制样品的研制和使用。
- 第 6 部分:标准样品包装通则。目的在于确立适用于标准样品包装和储存运输方面的基本规则。
- 第 7 部分:标准样品生产者能力的通用要求。目的在于规定有关组织管理和运行要求,以确保标准样品生产者能够生产出符合要求的标准样品。
- 第 8 部分:标准样品的使用。目的在于为标准样品的正确使用提供指导。

本文件解释了如何研发获得标准样品可靠特性值的各种方法,使得特性值能够溯源至合适的特定标准。

随着 RM 生产方法的发展,RM 的种类也随着技术的进步而增加,这就增加了对 RM 生产中更广泛适用的技术指导的需求。同时,各实验室越来越多地使用 ISO/IEC 17025 和 ISO 15189,需要对有证标准样品的计量溯源性有更加清晰的说明。

本文件在 GB/T 15000.3—2008(ISO 指南 35:2006,IDT)的基础上结合国情增加脚注,提供了更广泛的均匀性、稳定性研究策略和指导,并对 RM 生产中计量溯源性的建立做出了具体规定。

标准样品工作导则

第3部分：标准样品

定值和均匀性与稳定性评估

1 范围

本文件对标准样品生产的下列有关概念进行了解释并提供了方法：

- 均匀性评估；
- 稳定性评估和有关特性稳定性的风险管理；
- 标准样品特性的定值和赋值；
- 标准值的不确定度评定；
- 标准值计量溯源性的建立。

本文件为 ISO 17034 的实施提供了支持。只要满足 ISO 17034 的要求，也可以使用其他方法。

本文件提供了对互换性评估需求的简要指导(6.11)，但未提供技术细节；简要介绍了定性特性的定值方法(9.6 至 9.10)，并提供了对此类标准样品均匀性检验抽样的简要指导(第7章)，但未涵盖定性标准样品均匀性和稳定性评估的统计方法。本文件也不适用于多元量，如光谱数据。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本(包括所有的修改单)适用于本文件。

ISO 3534-2 统计学词汇及符号 第2部分：应用统计(Statistics—Vocabulary and symbols—Part 2: Applied statistics)

注：GB/T 3358.2—2009 统计学词汇及符号 第2部分：应用统计(ISO 3534-2:2006, IDT)

ISO 3534-3 统计学词汇及符号 第3部分：试验设计(Statistics—Vocabulary and symbols—Part 3: Design of experiments)

注：GB/T 3358.3—2009 统计学词汇及符号 第3部分：试验设计(ISO 3534-3:1999, IDT)

ISO Guide 30 标准样品 常用术语及定义(Reference materials—Selected terms and definitions)

注：GB/T 15000.2—2019 标准样品工作导则 第2部分：常用术语及定义(ISO Guide 30:2015, IDT)

ISO/IEC Guide 99 国际计量学词汇 基本和通用概念及相关术语(VIM)[International vocabulary of metrology—Basic and general concepts and associated terms (VIM)]

注：本文件将“国际计量学词汇”用“VIM”表示。

3 术语和定义

ISO Guide 30、ISO/IEC Guide 99、ISO 3534-2、ISO 3534-3 中界定的以及下列术语和定义适用于本文件。同一个术语有多个定义时，优先采用 ISO Guide 30 的定义。

ISO 和 IEC 维护的用于标准化的术语数据库地址如下：

- ISO 在线浏览平台：<http://www.iso.org/obp>

——IEC 电子开放平台: <http://www.electropedia.org/>

3.1

标准样品 reference material; RM

具有一种或多种规定特性足够均匀且稳定的材料,已被确定其符合测量过程的预期用途。

注 1: 标准样品¹⁾是一个通用术语。

注 2: 特性可以是定量的或定性的(如物质或物种的身份)。

注 3: 用途可包括测量系统的校准、测量程序的评估、给其他材料赋值和质量控制。

注 4: ISO/IEC Guide 99:2007^[3]有类似的定义[5.13],但限定“测量”术语仅用于定量的值。然而,ISO/IEC Guide 99:2007 的[5.13]的注 3 中明确包括定性特性,称作“标称特性”。

[来源:GB/T 15000.2—2019,2.1.1]

3.2

有证标准样品 certified reference material; CRM

采用计量学上有效程序测定的一种或多种规定特性的标准样品,并附有证书提供规定特性值及其不确定度和计量溯源性的陈述。

注 1: 值的概念包括标称特性或定性属性,如身份或序列,该特性的不确定度可用概率或置信水平表示。

注 2: 标准样品生产和认定所采用的计量学上的有效程序已在 ISO 17034 和 ISO Guide 35 中给出。

注 3: ISO Guide 31^[2]给出了证书内容的编写要求。

注 4: ISO/IEC Guide 99:2007^[3][5.14]有类似的定义。

[来源:GB/T 15000.2—2019,2.1.2]

3.3

测量模型 measurement model

测量中已知涉及的所有量之间的数学关系。

[来源:ISO/IEC Guide 99:2007,2.48^[3]]

3.4

特性值 property value

〈标准样品的特性值〉表示标准样品的物理、化学或生物所对应特性量的值。

[来源:GB/T 15000.2—2019,2.2.1]

3.5

标准值 certified value

赋予标准样品特性的值,其具有不确定度和计量溯源性陈述,并在标准样品证书中标明。

[来源:GB/T 15000.2—2019,2.2.3]

3.6

信息值 indicative value; information value; informative value

标准样品的量或特性的值,其仅用于提供信息。

注: 信息值不能作为计量溯源链的参考标准。

[来源:GB/T 15000.2—2019,2.2.4]

3.7

校准样品 calibrant

用于设备或测量程序校准的标准样品。

[来源:GB/T 15000.2—2019,2.1.21]

3.8

质量控制样品 quality control material

用于测量质量控制的标准样品。

1) Reference Material 也译为参考物质、标准物质。

[来源:GB/T 15000.2—2019,2.1.22]

3.9

同步稳定性研究 **isochronous stability study**

在短时间内,对暴露于不同储存条件和时间的标准样品单元,所进行的稳定性试验研究。

3.10

生产 **production**

〈标准样品的生产〉为发布和维护(有证或非有证)标准样品开展的所有必要活动和任务。

注:活动包括策划、控制、材料处置与贮存、材料加工、均匀性和稳定性评估、定值、特性值及其不确定度的赋予、标准样品证书或其他文件的批准和发布。

[来源:GB/T 15000.2—2019,2.3.7]

4 符号

下列符号适用于本文件。

a	均匀性研究中标准样品单元数
d	测量偏倚
k	包含因子或(用作下标)编号
L_d	按照 ISO 11843-1 ^[48] 方法计算的检出限(净状态变量的最小可检出值)
$N_{\min}$	批量高于 100 个单元的均匀性研究中 RM 的最小单元数
N_{prod}	在单一批中生产的 RM 单元数
n_r	区组或嵌套均匀性研究设计中的轮次数
p	实验室间认定演练中实验室均值数
s_{bb}	均匀性研究中单元间方差分量,用标准差表示
s_r	重复性标准差
s_R	再现性标准差
t_{its}	长期稳定性研究的持续时间
U_{CRM}	CRM 特性值的扩展不确定度
u_{bb}	与单元间变异性相关的标准不确定度
u_{char}	与定值研究中所赋的值相关的标准不确定度
u_{CRM}	CRM 特性值的标准不确定度
u_{trg}	特性值的目标测量不确定度,用标准不确定度表示
u_{hom}	与非均匀性相关的标准不确定度
u_{its}	与长期稳定性相关的标准不确定度
u_{mon}	监测时与测量 RM 所得值相关的标准不确定度
u_{tm}	与材料运输稳定性相关的标准不确定度
u_{wb}	与单元内非均匀性相关的标准不确定度
x_{CRM}	CRM 的特性值
$\hat{x}$	由稳健统计量获得的估计值
x_{mon}	监测时由测量 RM 特性值得到的值
x	物质的量分数
y_{char}	在定值研究中赋予标准样品的值

在特定条款中使用的其他符号在本文件第一次出现时定义。

5 约定

本文件采用如下约定。

- a) 某个规定的被测量具有唯一的“真值”。
- b) 除另有说明,本文件描述的所有概率评估均假设为正态分布。
- c) 本文件中,使用不确定度传播律合成测量不确定度的贡献。也可使用测量不确定度评定的其他方法,在某些情况下也有此必要。关于这些问题的进一步指导见 ISO/IEC Guide 98-3《测量不确定度 第3部分:测量不确定度的表达指南(GUM:1995)》及其补编(见参考文献[5]和[6])。

注1:与不均匀性相关的单元间变异和因不稳定性所引起的变化可能不是正态分布,并可能导致不对称分布。

注2:本文件“GUM”用以指代“测量不确定度的表达指南”。

6 标准样品生产概述

6.1 概述

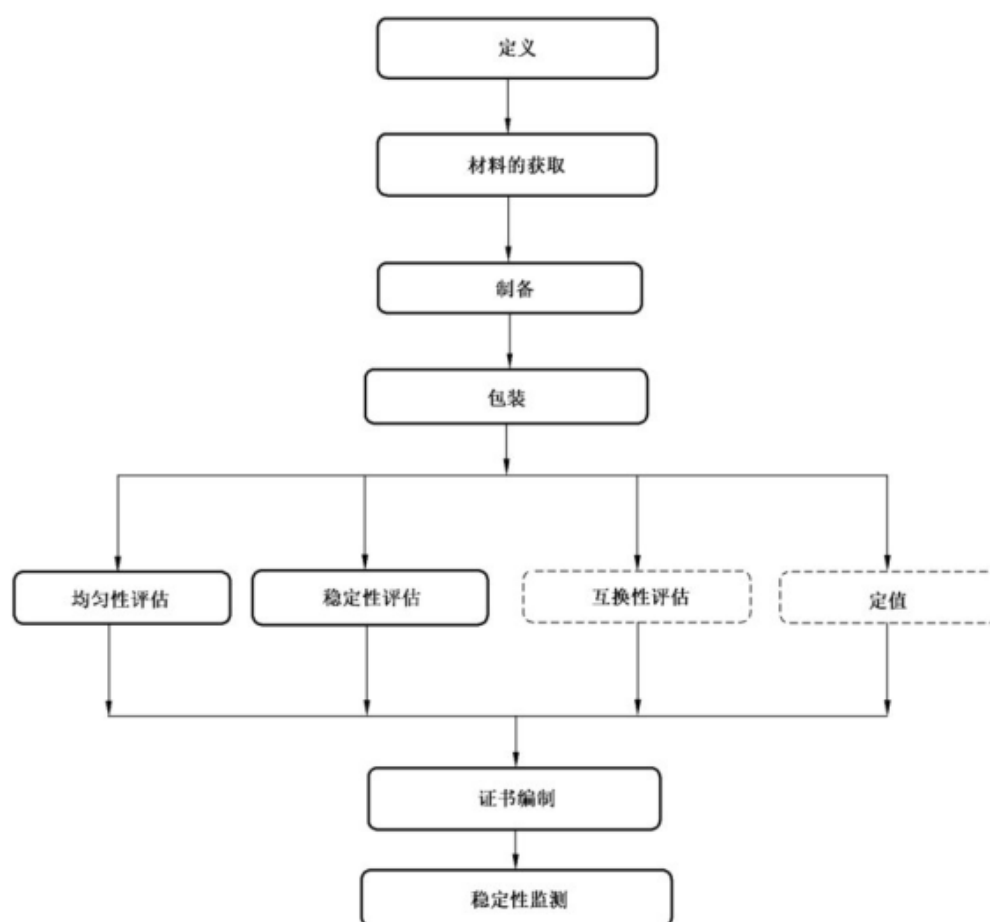
在开展 RM 生产和分发的任何实际活动之前需要进行精心策划。以下各部分概述了标准样品生产中所涉及的步骤,并对每个步骤中涉及的主要问题进行了描述。有关均匀性评估、稳定性评估和定值的指导详见第7章、第8章和第9章。

6.2 项目设计概要

标准样品生产涉及以下步骤:

- a) 定义 RM,如基体、拟定值的特性及预期水平、材料的预期用途、目标不确定度^[72](仅适用于 CRM);
- b) 策划获取原材料的程序;
- c) 策划标准样品制备加工的程序;
- d) 选择适合定值、均匀性和稳定性研究的测量程序;
- e) 考虑每个被测特性的计量溯源性,特别是 CRM,需要计量溯源性陈述;
- f) 评估均匀性;
- g) 评估稳定性;
- h) 评估互换性(如需要);
- i) 标准样品定值;
- j) 合并来自均匀性、稳定性研究的结果,对于 CRM,评定标准值的测量不确定度;
- k) 编制证书或者产品说明书,如合适,编制研制报告;
- l) 规定储存和运输条件;
- m) 监测生产后稳定性。

主要阶段见图1。



注 1: 本图为生产和维护标准样品的主要步骤示意图,虚线部分是可选步骤。

注 2: 图中“包装”包括分装至单个单元(用于分发的合适容器)中。

注 3: 图中“证书编制”包括可与标准样品一起提供的所有类型文件,如证书、产品说明书和研制报告等。

图 1 标准样品生产示意图

6.3 原材料的获取

在 RM 生产项目中,第一项任务是要获取足够数量具有所需特性的原材料。具有特定特性材料的生产在 9.3.4 中简要说明。所需材料的数量取决于以下因素:

- 在 RM 预期寿命中需要分发的 RM 单元数;
- 均匀性研究所需单元数;
- 稳定性研究所需单元数;
- 候选 RM 定值所需单元数;
- 在 RM 预期寿命中稳定性监测所需单元数;
- 每个 RM 单元所计划的量,至少须满足一次测量;
- 一项或多项可行性研究的需要;
- (可选)应对意外情况所需单元数,如回应客户问询的跟进研究,在储存条件下因发生显著变化未来需要重新认定,或增加认定特性的数量。

需要分发的候选 RM 单元数通常至少在某种程度上是一个商业问题,在委托采集和加工前宜仔细加以考虑。另外,材料储存时预期的长期稳定性也会影响生产材料的实际数量。一种谨慎的做法是控制不太稳定的材料生产的单元数量,这样可以防止因随时间发生不可避免降解带来的损失。

6.4 可行性研究

可行性研究是短期研究,旨在解决对足够均匀和稳定 RM 的生产和定值可行性的担忧。例如,在项目初期,通过小规模可行性研究,可以回答诸如制备 RM 的最佳方式或确保材料足够稳定等问题^[7]。

如果预期使用实验室间研究方式进行定值,可行性研究可识别可能的误差来源,而且可以使定值参加者优化其设备和程序。

注:在实验室间定值(见第 9 章)演练活动中,如欲通过可行性研究测试或改善参加者的能力,使用候选 RM 以外的材料,可以避免因对候选 RM 的提前了解,造成参加者结果的不适当偏倚。

6.5 标准样品加工

加工可能涉及一系列过程,如包括:

- 人工标准样品的合成、制造或者配制;
- 天然材料的干燥、冻干、研磨和/或过滤;
- 稳定剂的添加;
- 包装前均匀化。

所使用的具体程序取决于特定材料,通常需要专家指导。

6.6 均匀性评估

对于所有 RM,均匀性都是一项重要要求,包括单元内均匀性和单元间均匀性两个部分。单元间均匀性对于确保每个 RM 单元的每个特性具有相同值至关重要,单元内均匀性在标准样品用户可以取子样测量时很重要。第 7 章给出了均匀性评估的详细指导。

6.7 稳定性评估

RM 要足够稳定,以满足其预期用途,使最终用户可以在证书有效期内任何时间点信赖赋值。通常,重要的是要考虑长期储存条件、运输条件和(如适用)RM 用户实验室储存条件下的稳定性。当 RM 允许重复使用时,还要考虑单元开启后的稳定性。第 8 章给出了稳定性评估的详细指导。

6.8 测量程序的选择

在标准样品生产项目中,每个需要测量的步骤可以使用不同的测量程序,这是因为定值通常要求最小偏倚的测量程序、并具有较小不确定度。均匀性研究主要要求可获得的最佳重复性,而经典稳定性研究通常要求测量程序在同一实验室内不同时间具有良好的精密度。关于均匀性研究、稳定性研究和定值所用测量程序的选择将分别在第 7 章、第 8 章和第 9 章讨论。

6.9 计量溯源性

计量溯源性是确保测量结果(包括用于标准样品定值的测量结果)在不同时空上具有可比性的关键。根据定义,CRM 附有每个标准值的计量溯源性陈述。由于 CRM 主要用来建立测量结果的计量溯源性,因此,正确选择规定的参考标准,建立特性值的计量溯源性,对 CRM 至关重要。建立计量溯源性将在 9.2 详细讨论。

6.10 定值和不确定度评定

定值是指作为生产过程的一部分,测定 RM 相关特性的特性值。RM 的定值将在第 9 章阐述。对于 CRM,标准值附有测量不确定度陈述,不确定度的评定将在第 10 章讨论。

6.11 互换性评估

RM 的互换性与 RM 的性能有关,是指由某一测量程序(通常为参考程序)定值的 RM 可作为另一测量或检测程序(适用于常规样品检测)的校准样品或质量控制样品(QCM)。当不同的测量程序对不同类型的检测材料有非常不同的响应时,这尤为重要。并非所有的 RM 都需要评估互换性,但对一些重要类别的 RM 却是必须的。

注:目前 ISO/REMCO 关于互换性评估^[9]的内容规定:

“标准样品生产者在下列情况下要进行互换性评估:

- a) 预期用途需要校准样品或者质量控制样品的互换性;和
- b) 标准样品生产者保证 RM 适用于预期用途。

注 1: 当预期用途包含生物测量中的校准或质量控制时,通常要求证明互换性;而当预期用途中不包含生物测量,且已知该程序对于标准样品和预期常规样品基体中的被测量具有足够的选择性时,则不需要证明互换性。

注 2: 当标准样品及其原料来源和处理方式与客户待测样品均相同时,如基体标准样品,则不需要建立互换性。”

6.12 运输事宜

几乎所有 RM 都必须运输到使用地点。生产后 RM 的运输的方式和条件与稳定性研究(见第 8 章)的需求有关,因此需要在项目早期就开始考虑运输条件。

注 1: 国家和/或国际运输法规可能限制了一些运输选择,禁止某些材料的运输,或者基于安全或其他原因规定了特殊的包装或预防措施。

注 2: 在海关或其他边境管制通关等官方程序所需时间,可能会增加某些目的地的装运时间。

6.13 赋值

赋值是指将均匀性和稳定性评估结果与定值研究的结果进行合成,确定特性值及其不确定度的过程,这些值随后在证书或者产品说明书上发布。

6.14 稳定性监测

大多数标准样品将会在 RM 生产者或经销商处长期储存。由于稳定性评估通常不能预测所有可能发生的变化,因此作为可能不稳定的风险管理的一部分,需要监测标准样品特性值的长期稳定性。对监测的要求在一定程度上取决于稳定性评估中获得的知识。有关稳定性监测的指导将在 8.10 中讨论。

6.15 重复批生产的标准样品

当根据既定程序重复生产标准样品时,可以降低对某些特性(特别是均匀性、稳定性和互换性)进行试验研究的需求。只要满足以下条件,凭借之前的经验是合理的:

- a) RM 的批生产过程未发生任何可能不利于最终使用的变更;
- b) RM 生产所用原料未发生任何不利于最终使用的变更;
- c) 前期使用同一工艺生产的 RM,无论在日常监测还是在用户使用中,均未发现可归咎于生产工艺的问题;
- d) 对标准样品的要求进行定期评审,评审时考虑标准样品的预期用途,确保生产过程符合目的。生产过程的一致性需要进行核查,例如,将在重复性条件下生产的连续批样品进行特性值的比较。

7 均匀性评估

7.1 概述

大多数 RM 是以“单元”(如瓶、小瓶或试片)成批制备的。所有分发单元的每个特性值在规定的

确定度范围内均相同非常重要,除非以单次使用单元销售,否则每个单元内的材料都是均匀的。因此,ISO 17034 要求对标准样品均匀性进行评估。

均匀性可以指标准样品各单元间特性值的变异,也可以指每个单元内的变异。评估单元间变异总是必要的。如果预期用途允许使用单元的一部分,如固体材料或液体材料的一小部分、或表面的一小块区域,评估材料单元内变异(单元内不均匀性)或提供控制单元内不均匀性影响的使用说明通常也是必要的。这些说明可包括如样品的重新混合、粉粒状材料的最小取样量,因为单元内不均匀性直接反映了代表整个单元的样品的最小量。

均匀性评估可包括使用该材料均匀性的之前证据(包括之前的试验证据)、对候选标准样品开展的均匀性试验研究、或两者均包括。多数情况下,试验研究是必要的。例外情况包括:很多高度均匀的材料,如之前试验研究已表明包装和贮存不影响均匀性的溶液或生产的标准样品每个单元有单独的赋值。关于需要试验研究的材料状况和类型,7.2 给出了进一步详细描述。

均匀性试验研究结果通常也用于定值模型中(见第 10 章)均匀性不确定度分量的计算,与其他不确定度分量相比,该不确定度分量大小可能会有很大的变化,这取决于 RM 及定值特性的性质。

为了进行均匀性研究,使用一个合适的抽样方案从批中选取一组单元,通常为 10 到 30 个,然后用适当的测量程序测量每个单元的特性值,并用合适的统计方法对结果进行评估以获得该材料单元间变异和单元内变异的信息。

为了获得可靠的结果,重要的是要:

- 选择拟研究的特性;
- 选取一组有代表性的单元;
- 选择具有足够重复性和选择性的适当测量程序;
- 按照适当的试验设计在适当条件下进行测量;和
- 采用有效的统计方法开展统计分析。

7.3 至 7.7 提供了每个步骤的指导,附录 C 中提供了计算的示例。

以前,均匀性研究是将单元间差异与测量精密度比较进行统计显著性检验,以确定某种材料是否均匀。本文件没有采纳该方法,而是把重点放在判定单元间标准差对于预期最终用途是否足够小。不过,统计显著性检验可以在 RM 生产中使用,如将不均匀性减小到与常规测量精密度比较不具有显著性,以决定是否需要进行进一步加工。但是,如果使用统计检验,均匀性试验宜能够检出重要的不均匀性,反过来又需要对测量程序的精密度、RM 单元数及每个单元重复次数进行充分结合。统计功效计算(7.4.2)能有助于确保足够有效的检验。

7.2 均匀性试验研究的需求

天然或具有复杂基体的材料,如食品、土壤、矿石和合金,其组成通常是不均匀的。尽管单元间差异的大小在均化后有时很小或甚至可忽略不计,但有时单元间差异与定值引起的不确定度相比还是大很多,因此由这种不均匀材料制备的 RM 宜进行均匀性试验研究。

制备纯化合物或纯化合物溶液的 RM(如定值的是纯度,而不是杂质)预期具有高度的均匀性,然而这些材料也会呈现某些不均匀性。例如,由于在加工和充装过程中产生密度梯度、局部污染和溶剂的蒸发、剩余溶剂容积的变化,以及含有包裹了不定数量气体的金属。此外,通常会认为这类材料标准值具有非常小的不确定度,甚至微小的不均匀性也可能很重要。因此,即使该材料在多数预期用途中预期是足够均匀的,还宜进行均匀性验证。验证可包括一个完整的均匀性研究或其他核查(如对纯有机物核查单元之间熔点的一致性)。

在下列情况下,材料的均匀性试验研究不是必需的:

- 当该材料是按照之前材料相同的程序重复生产,且之前材料已经过试验表明足够均匀;或
- 生产过程已经过确认,由此证明能够持续生产足够均匀的批量材料。

有时以这种方式生产的标准样品,如乙醇校准溶液或元素校准溶液,这些溶液通过质量比配制并充分混合,以确保混合物足够均匀、符合预期用途。

如果均匀性的保证依赖于经过确认的生产过程,则宜使用质量控制程序来证实生产过程的一致运作。这类程序可包括:绘制被测单元的极差控制图或标准差控制图,用于监测少量被测单元的极差或标准差,或确定每次定值中值的极差评价标准。

注 1: 6.15 提供了依赖之前生产批获得经验的附加指导。

注 2: ISO 7870-2^[73] 给出了使用极差控制图和标准差控制图的指导。

7.3 拟研究的特性

如果要通过试验确定均匀性,通常有必要测定每个感兴趣特性的均匀性,也即对于预期用途,材料标称的每个特性都是足够均匀的。

在下列情况下,可通过检查较少数量的选定特性评估所有感兴趣特性的均匀性:

- 此类材料中特定特性值之间有足够高的统计相关性,可以由一个或多个其他特性值对另一个特性值进行有用的预测;和
- 能够证明特定的一组特性是充分密切相关的(例如,由于它们多以混合物的某个特定组分存在),对此组中一个特性的测量为同组中其他特性的均匀性提供了证据。

基于已建立的化学或物理关系适当选取的特性组能代表一个更大特性组的均匀性,这是很重要的。例如,RM 矿物相中元素间伴生将支持以下假设,即 RM 未选取的元素也具有相似程度的均匀性。

在使用有关特性组进行均匀性试验评估时,对未进行试验研究的特性,宜取得均匀性的附加证据。证据宜足以证明:那些没有进行试验研究的特性,其与不均匀性相关的不确定度没有被低估。

注:例如,此类证据可以从有关材料的文献中,或通过材料稳定性研究或定值获得。

7.4 统计有效的抽样计划

7.4.1 均匀性研究的最少单元数量

7.4.1.1 定量特性的均匀性研究

通常定量特性的均匀性研究旨在提供因不均匀性引起的方差和加工过程产生的任何(可能是非线性)趋势的信息。为实现这些目标,单元数量要足以对单元间方差做出合理估计,并选取足够的单元以便清楚地了解存在的任何趋势。

根据现行做法,当自由度为 9 或更大时,利用单元间方差估计不确定度是可接受的。对于均匀性研究的简单方案或随机区组设计(见下文),这相当于至少选取 10 个单元。如果采用嵌套设计进行多轮测量时,在每轮测量的一个单元组中宜包括额外的单元,以保证所需的自由度。7.6.4 将进一步详细说明嵌套设计所需额外单元数。

加工过程产生的趋势经常表现为初期产生趋势后期稳定输出的模式,或者初期稳定输出后期产生趋势,或者两者的结合。如果出现这些特征,通过丢弃加工开始或结束阶段受影响的单元,通常可以提供均匀的材料。然而,只检查 10 个单元可能无法提供足够的信息说明在漫长加工过程快结束时出现的趋势。因此,当前最好的方法是随着生产的单元总数 N_{prod} 的增加,相应增加检查单元数。典型的建议是在 $\sqrt[3]{N_{\text{prod}}}$ 和 $3 \times \sqrt[3]{N_{\text{prod}}}$ ^[74] 之间。结合上述自由度要求,对于进行定量特性定值的材料均匀性研究推荐的最小单元数 N_{min} 为:

$$N_{\text{min}} = \max(10, \sqrt[3]{N_{\text{prod}}}) \quad \dots\dots\dots (1)$$

式中, $\max(\dots, \dots)$ 表示括号内的最大项。

注 1: 对定量特性标准样品的定值,通常不需要检查 30 个以上单元。

注 2: 当生产批批量为 100 个或更少时,7.4.1.3 给出了确定最小单元数的进一步指导。

示例:假设标准样品生产者制备了 3 000 个单元候选 RM,准备进行单一轮均匀性研究,那么 $N_{\text{prod}} = 3\ 000$ 和 $N_{\text{min}} = \max(10, \sqrt[3]{N_{\text{prod}}})$,结果是 14.4,因此该研究需要 15 个单元进行均匀性研究。

7.4.1.2 定性(“标称”)特性的均匀性研究

以定性或“标称”特性定值的标准样品,均匀性研究选择的单元数宜根据 ISO 2859 系列标准^[8]给出的计数抽样检验程序或类似的指南进行设定。

如果仅通过抽样检出低比例的缺陷单元,则计数抽样检验程序将导致非常高的检验数量。例如,若要以高置信度检验批不合格率是否为 1%,需要检验大约 300 个随机抽取的单元,根据检出的不合格单元数进行判定(假定批量远大于 300,样本量由批不合格率为 1%,置信水平为 95%时的统计功效计算得到)。对于通常情况下的标准样品批量,该方法是不可行的。因此,对于定性材料,均匀性试验研究可能仅限于检验非预期的总不均匀性(如不合格率大于 10%),并综合有关材料来源、加工过程等其他信息作为均匀性判定的证明。

如果对定性特性的定值是对该材料的每个单元进行单独验证的,则无需对该特性进行均匀性检验。

7.4.1.3 小生产批

一些标准样品以 50 个或更少的单元进行小批量生产,例如,次级气体校准标准。对于此类小生产批,7.4.1.1 中规定的最少单元数通常相当于可用单元的很大部分。如果批量少于 100 个单元,宜从批中随机选取三个或批大小 10%之间的较大的单元数进行均匀性评估。重复测定数宜在实际条件允许的情况下尽可能高,以便对使用的单元数提供最佳的可用检验功效。功效分析(7.4.2)可能有助于考虑所需的重复测定水平。例如,选用 3 个单元,对每个单元进行 4 次观测与使用 10 个单元,对每个单元进行 2 次观测在检出单元间正态分布偏差时具有大致相同的检出功效,但在其他均匀批中检出偶然缺陷单元的概率要低得多。

7.4.2 统计功效分析的使用

统计功效分析可用于协助选择合适的单元数和重复次数进行均匀性研究。功效分析用于控制统计检验无法检出某个不均匀特性显著不均匀的概率。功效分析是一项需要专门研究的课题,目前可用的软件很多,其中一些可免费使用。因此,最常见的功效分析示例是基于单因子方差分析(见 7.7.3)可以检验单元间是否有显著差异的假定,计算检测单元数和重复测量数。

统计功效分析在比较不同的建议策略时是有用的,但在使用时要相当小心。尤其是,重复次数和单元数的选择过分依赖于样本间差异的分布假设。例如,设单元(真)平均值服从正态分布,功效计算软件中常见的缺省值会推荐一个大的重复次数和一个小的单元数。一种最有可能的不均匀模式是在一个很大的均匀总体中有很小比例的不一致单元,这不是一种好的选择。因此,当单元(真)平均值的分布信息未知时,功效分析在确定了建议的单元数之后协助选择重复次数是最有用的。

注 1: 使用功效分析时,通常选择 95%的显著性水平和 80%的预期功效,但请注意,它的功效实际大于 7.4.1.1 所考虑的最少单元数和重复次数。

注 2: 参考文献[57]给出了适用于标准样品研究的一些简单功效计算的细节。

7.4.3 均匀性研究的抽样策略

均匀性研究中选取单元(样品,瓶)的抽样方案通常是简单随机抽样、分层随机抽样或系统抽样中之一。用于均匀性研究标准样品抽样的一般程序如下。

- 简单随机抽样,从批中选取一个随机样本。
- 分层随机抽样,通常将批分成大小相同的若干部分(通常按照生产或包装顺序,有时按照位置),并从每个部分随机选取相等数量的单元(通常为一个)。

——系统抽样,将被抽样单元按一定顺序排序,设置一个间隔 n_{syst} ,并在 1 和 n_{syst} 之间设置一个随机起始点 n_1 ,然后在 n_1 、 $n_1 + n_{\text{syst}}$ 、 $n_1 + 2n_{\text{syst}}$ 等处选取单元,所选 n_1 和 n_{syst} 完全覆盖生产批和所需的单元数。

确定随机样本宜使用随机数生成软件或随机数表,手动随意挑选不等同于随机抽样。

当单元没有已知的排序时,或者大部分单元用于均匀性研究的生产批较小,简单随机抽样是最合适的。对于大多数情况,建议使用分层随机抽样,因为这保证了均匀性研究所选取的单元在整个批中大致均匀分布,避免抽样位置与制备过程中任何循环效应之间的重合。当批中重复效应或趋势有风险时,可以选用系统抽样。

抽样方案宜考虑制备和储存样品方法的潜在缺点(例如,包括加工和充装的趋势以及分装之前储存时可能的沉降),从而对已制备批进行严格检查。

单元通常宜在抽样前进行编号、或在抽样后以加工顺序编号,以便后续的趋势分析。

其他的抽样方案也可以使用,只要能证明所得到的方差估计不会因为所选的抽样方案而产生偏差。

7.5 均匀性研究的测量程序选择和实施

7.5.1 测量程序的选择

用于进行均匀性研究的测量程序宜主要选择在每轮测量预期的时间内具有良好精密度(即良好的重复性标准差 s_r)的,且如果单元要在几轮中进行随机测量,则要选择具有良好轮间精密度的方法。测量标准差宜小于每个被测特性值预期的不确定度。理想情况下,均匀性研究测量程序的重复性标准差宜小于所要求的标准不确定度的三分之一,即如果特性值的目标测量不确定度是 u_{trg} (以标准不确定度表示),则该程序的重复性标准差在理想情况下宜满足:

$$\frac{s_r}{\sqrt{n_{\text{al}}}} \leq \frac{u_{\text{trg}}}{3} \quad \dots\dots\dots (2)$$

式中, n_{al} 为从供测量的每个单元中取出 n 等份后对每份进行观测的次数。

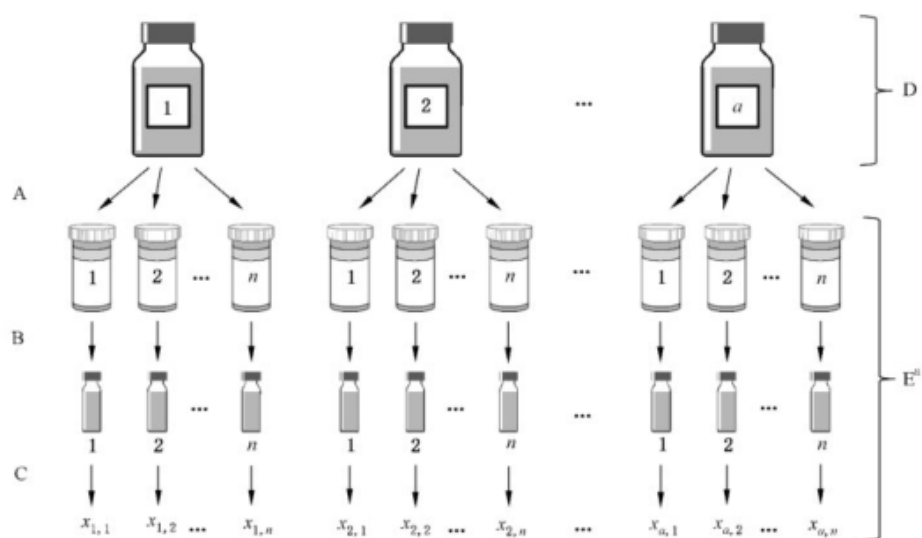
如果在实践中不能满足该精密度要求,例如 RM 单元不允许大量重复测试,则尽力确保 $s_r/\sqrt{n_{\text{al}}}$ 不超过目标标准不确定度 u_{trg} ,并且宜考虑增加单元数。

注 1: 除非重点是为了发现不常见的、高度不一致的单元,增加单元数对于克服精密度差的影响不如增加每个单元的观测次数或提高测量精密度更为有效。统计功效计算(见 7.4.2)可用于选择合适单元数和重复次数进行试验设计,以在满足公式(2)中精密度要求时获得类似的保证。

注 2: 选择测量程序的另一个考虑是将每个样品转变为可测量形式所需的样品制备数量和复杂程度。很少需要或不需要样品制备的程序(所谓的无损方法)比需要复杂、多步骤样品制备的程序具有更好的精密度。

7.5.2 均匀性研究的测量实施

如果可能,将不同 RM 单元结果的离散性与重复性条件下(7.6.1)的离散性进行比较非常重要。“重复性条件”表示完整测量程序的重复。因此,只要可能,将完整的样品制备(“转化”)和测量步骤应用于从每个单元取得的每个子样或试样(见图 2)是非常重要的。否则,重复观测之前样品制备的任何变异都将会增加单元间不均匀性(见图 3),导致过高估计与不均匀性相关的不确定度。



标引符号说明：

A —— 取子样；

B —— 制备；

C —— 测量；

D —— 用于观测到单元间变异；

E —— 用于观测到单元内变异的^a操作。

^a x_{ij} 代表单元 i 的第 j 试样。

图2 单元间均匀性研究的设计图

图2是一种理想的情况,其中由标准样品单元取子样是可能的,并从该批的每个样品中取出了多份(n)试样,然后单独进行制备(转化)用于测量。在这种情况下,单元内方差包括由取子样、制备和测量所产生的变异。从获得单元间标准差的无偏估计来看,这是理想的情况,因为这些影响都不会对所估计的单元间方差有贡献。

图3为不能从样品中取子样或只是没有取子样的情况。从 a 个标准样品单元中每个单元中仅取出一个子样制备用于测量,并对所制备的每个试样进行 n 次测量。在该设计中,样品制备和任何取子样的效应都包括在观测到的由试验计算的单元间方差中,所计算的单元内方差仅包括测量的重复性。这个程序将导致更大的单元间方差估计,从而高估了单元间的不均匀性。

因此,开展均匀性研究时,只要可能,在每个子样上均要完成测量程序的所有步骤。

7.6 均匀性研究设计

7.6.1 均匀性研究的目的

均匀性研究使用统计方法对RM几个单元观测值的离散性与测量程序的精密度进行比较,以便确定单元间标准差,然后用于计算与不均匀性相关的不确定度。这通常需要对每个RM单元进行相同次数的重复测量,见图2。均匀性研究中将标准样品单元测量结果的离散性与重复性条件下结果的离散性(参见参考文献[11])进行比较,以获得单元间项的最佳估计。这可以通过以下设计实现:

- 在单一个测量轮次中对指定特性进行所有的均匀性测量;或
- 使用试验设计可以分别估计测量轮内、测量轮间和单元间方差,为此最有用的设计是简单随机区组设计和平衡嵌套设计。

所考虑的三种设计如图4所示。使用这些不同设计的典型试验在7.6.2至7.6.4中有更详细的描述。数据分析分别包含在7.7.3至7.7.5中。

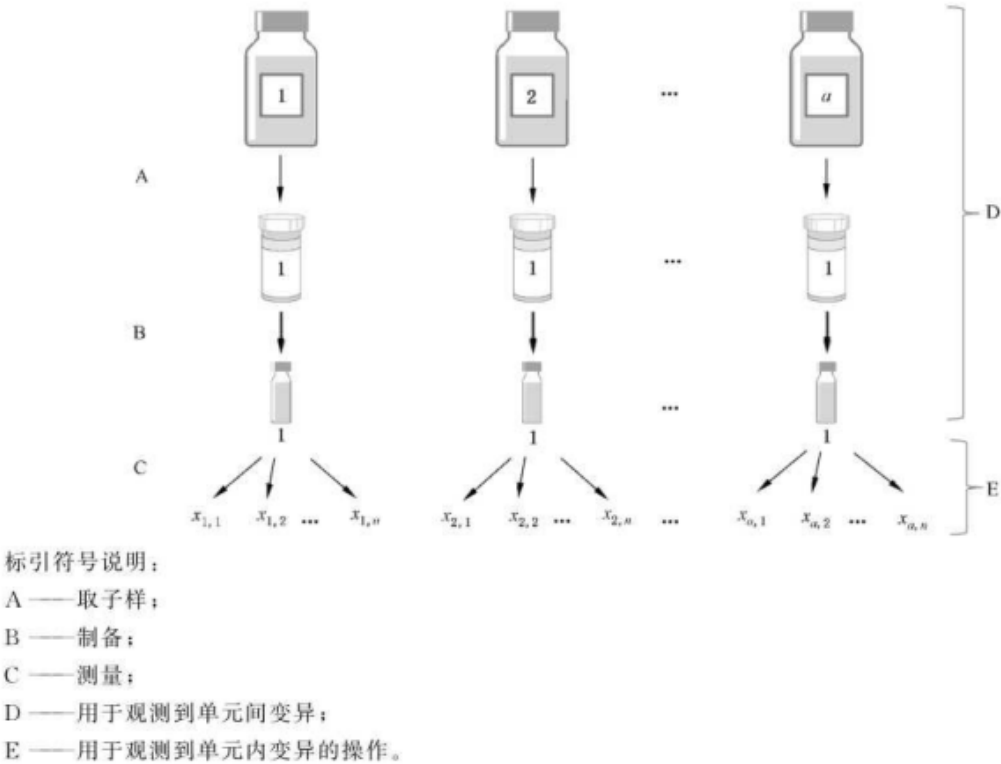


图3 单元间均匀性研究的备选设计图

a)	第1轮:	c	a	b	a	f	e	b	g	j	d	h	k	c	h	i	j	k	f	l	d	i	l	g	e
b)	第1轮:	a		e	j		b		i	d		h		c		f		l		g		k			
	第2轮:	g		d		j		f		e		i		k		l		h		b		a		c	
c)	第1轮:	c		a				a		h		b				h		b				b		c	
	第2轮:	k				i				i		l			e		e				l		k		
	第3轮:	g				j				d				f		g				f		j			

标引符号说明：

a——简单随机设计：所有单元在单轮中按照随机顺序重复观测两次。

b——随机区组设计：每轮对每个单元进行一次观测，每轮分别随机化。

c——嵌套设计：4个单元随机分配至3轮中，每轮中每个单元按照随机顺序重复观测2次。来自相同单元的观测偶尔是相邻的；见7.6.2注1。

所有图解均展示了从“a”到“l”命名的12个RM单元重复两次观测的设计。

图4 均匀性研究设计图

7.6.2 基本的均匀性研究设计——单轮测量

如果测量系统在所需观测的持续时间内是稳定的，所有观测值均可以在单个测量轮中获得，则可以使用单因子方差分析设计（见7.7.3、B.1和C.1）评估单元内和单元间标准差。测量宜以这样的方式进行：

- i) 测量系统的趋势不会被误解为RM单元间的差异；和
- ii) 如果需要，测量趋势可以与样品的批趋势分离。

最佳实现方式是对用于均匀性研究的样品以完全随机顺序进行重复测量。图4 a)是这种设计的说明性示例。

RM 生产者也可以使用系统重复观测安排,从中可以看出:

- a) 测量误差的合理期望模式,包括与轮内趋势和步骤的响应变化,不会引起对单元间方差估计的明显偏差;
- b) 后续检查和数据分析程序能从均匀性研究期间测量趋势中区分出加工过程(包括包装)趋势。

随机选择的顺序可能会偶然导致单元几乎按照加工顺序排列或者同一 RM 单元的几乎所有重复测量在测量序列中紧密地连在一起。若所建议的随机排序与加工序列产生明显强相关性,或使得大多数或所有单元的重复测量紧密成群,则该随机排序宜被拒绝,并选择新的随机排序。

注 1: 随机化的顺序不能保证来自同一 RM 单元的各个重复测量不相邻。在完全随机的设计中,偶尔出现相邻重复测量是常见的和可接受的。

注 2: 以下示例说明了在仪器响应随时间呈简单线性趋势的情况下,一个系统的正反向观测排序不会使单元间项引起明显偏倚^[12]。

示例: 假设有 10 个样品用于均匀性研究,每个重复测量 3 次,那么,可能的测量方案如下:

第一次重复: 1 - 3 - 5 - 7 - 9 - 2 - 4 - 6 - 8 - 10

第二次重复: 10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

第三次重复: 2 - 4 - 6 - 8 - 10 - 1 - 3 - 5 - 7 - 9

7.6.3 随机区组设计

在 a 个标准样品单元、每个单元进行 n_0 次重复测量的随机区组设计中,最简单的随机区组设计包含 n_0 轮测量,并且每轮中每个单元按随机顺序各测量一次。各轮测量宜单独随机化。图 4 b) 是这种设计的说明性示例。无重复试验的双因子方差分析或其他方法(见 7.7.4 和 7.7.6),可不受轮次影响,估计单元内和单元间标准差。

如果在单个测量轮中因时间限制或仪器限制所有单元不能包含所需重复测量次数,随机区组设计是合适的。当测量过程易在较长轮末端漂移时,它也是有用的。在这种情况下,用随机区组设计的几个短轮比单一轮可获得更好的精密度(作为单元内标准差来测量)。

注 1: 在经典统计分析(使用平方和计算的单因子方差分析)中,对于简单随机区组设计,单元间标准差的自由度等于 $a-1$,其中 a 是单元数。不过,与单元内方差相关的剩余自由度每增加一个测量轮就减少一个。

注 2: 在如上所述的简单随机区组设计中,所计算的单元内标准差包括由于测量程序而引起的变异和由于任何单元内不均匀性引起的变异。

7.6.4 平衡嵌套设计

在 a 个单元的平衡嵌套设计中,进行了 n_r 轮测量,每个测量轮包括相同的单元数 a_w ,每个单元都重复测量 n_0 次,所有的重复测量均在相同的测量轮中进行,每个测量轮是单独随机的。图 4 c) 是一个说明性的例子。

嵌套设计减少了与单元间标准差相关的自由度。对于这些设计,要增加测量的单元数以保持足够的自由度。这通常需要在嵌套设计中,对每个附加的测量轮添加至少一个额外的单元。此外,如果要使用的统计方法依赖于平衡设计,则宜进一步增加足够的 RM 单元,以确保每个测量轮都包含相同数量的 RM 单元。

注: 平衡设计简化了基于平方和的经典单因子方差分析。如果要手工或通过软件(例如,提供基本单因子方差分析功能的电子表格应用程序)进行方差分析,则保持每组相同数量的单元是有利的。新的方差估计方法(见 7.7.6)在计算上更复杂,但受不平衡设计的影响较小,因为在每组中出现不同数量的 RM 单元或测量重复次数。

示例: 某个标准样品生产者使用的测量程序在单轮测量中能获得多达 12 个观测值,制备 2 000 个候选 RM 单元。假如在单轮测量中完成研究,由公式(1)得到 $N_{\min} = 12.6$ 或 13 个单元。由于标准样品生产者在单轮测量中不能获得对 13 个单元的重复观测值,因此,标准样品生产者选择使用了 3 个测量轮的嵌套设计,这需要增加两个测量轮,单元数相应地从 13 个增加到 15 个。这样,标准样品生产者分 3 组(每组 5 个单元),共测量 15 个单元。

7.6.5 可替代策略

并非总是可以执行上述所列设计之一。例如,标准样品单元可能不允许再分割进行重复分析,可能因为所提供的样品量仅够用于单次测量(因为单元一旦打开后就不够稳定了),或是因为单元是不可分割的(如单个的夏比试片)。其他情形包括非常短的测量轮(例如,每次仪器运行仅得到 3 个观测值)或需要考虑轮影响之外的其他变异来源。

如果单元无法(再)分割,最常见的策略是每个单元测量一次,并将这些测量的标准差(或更常见的方差)与测量重复性的独立估计进行比较。例如,可以将 10 个单元单次测量的方差与某个未分装前、仔细均化制备的同一材料 10 次观测的方差进行比较,或者与基于方法确认研究的良好精密度估计进行比较。然后将方差的差异作为单元间方差的估计值。

如果可用的测量跨度太短,不允许进行完全随机区组或嵌套设计,采取的策略可以包括,例如:

- 在短的测量轮中随机分配重复测量次数,然后进行基本设计的统计分析,这可以增加单元内计算的方差,因为轮的影响就被包括在组内项中;
- 使用“平衡不完全区组”设计,其中许多短轮中的每一轮获得一组仔细选择的重复次数;平衡不完全区组设计超出了本文件的范围,但在有关试验设计的参考文献[11]中有描述了;统计分析可使用经典或其他方法。

如果要控制多个变异来源,可以使用其他可获得的设计,例如重复拉丁方设计^[12]或其他结构化试验设计^[11]。

7.7 评定均匀性研究

7.7.1 测量趋势和离群值的初步检查

7.7.1.1 测量趋势

为了解测量过程中的趋势,宜首先对试验数据(包括任何质量控制观测值)进行检查。最有效的做法是目视检查按试验轮顺序绘制的数据散点图。回归分析可用于检验统计上显著的趋势,例如,如果近似线性趋势是明显的,可使用简单线性回归模型。

在一个适当随机化的均匀性试验中,测量趋势通常会扩大单元内方差,从而可能掩盖重要的单元间影响。如果测量结果的显著趋势被证实,宜采取措施减小其对后续分析的影响。典型措施包括:

- 剔除受影响的测量轮并在消除趋势来源后进行重复测量;
- 在进行后续方差分析之前,基于一个适宜的平滑函数(如线性回归)进行校正,对趋势的预先校正减小了剩余自由度,这宜在后续方差估计中加以考虑;
- 在数据分析中包括趋势,尽管需要专门的统计软件和适当的统计专业知识,但该程序可根据自由度的任何损失进行相应调整。

注 1: 这些选项的示例见附录 C 中示例 C.4。

注 2: RM 生产者要制定评价趋势是否显著的标准。在没有其他评价标准的情况下,置信水平为 95% 的统计显著性作为均匀性研究中趋势判定的指标。

注 3: 在某些情况下,例如在具有三轮或更多轮的嵌套或随机区组设计中,如果剔除单个有问题的轮后,剩余轮中的精密度足够好,能保留足够的数据用于有效的均匀性评估,则无需重复有问题的轮。

7.7.1.2 基本均匀性研究中的离群值

许多测量结果是作为均匀性研究的一部分而产生的,但由于仪器或其他可归咎的原因,一些技术问题可能导致偶然的离群结果,因此,宜对离群的原因进行调查并采取适当的纠正措施。相应的纠正措施可包括,例如:

- 剔除单一离群观测值或剔除显示此类观测值的某个单元的所有结果(见下面的注 2)；
- 在短时间内重新测量某个可疑单元或可疑子样；
- 重复被几个离群值影响的完整测量轮或完整试验。

根据 RM 单元的某个极端均值拒绝该标准样品单元的均匀性研究结果是不合适的。

注 1：离群结果可以通过目视检查或通过使用离群值检验如格拉布斯(Grubbs)检验加以识别。ISO 5725-2^[76]给出了格拉布斯检验。ISO 16269-4^[76]给出了图形离群值检验和其他多个离群值检验。

注 2：单元内离群值可能是由单元内不均匀性引起的。如果可能存在单元内不均匀性，在舍弃任何离群的重复结果之前，谨慎的做法是做进一步调查。

注 3：从某个单元中剔除某个离群值将导致不平衡。如果所用统计方法依赖于平衡设计，只要该单元的非离群值与数据组的其余部分一致，并且在研究中保留的单元数是足够的，则可以删除该单元的数据。

注 4：由于仪器的影响，比较大的均匀性研究(例如，对于某个给定的特性，研究涉及 20 个或更多个 RM 单元)自然会产生一个以上单元内离群值，因此，在调查了离群值的原因之后，舍弃一个以上此类观测值可能是合适的。

7.7.2 加工趋势的检查

按照加工顺序绘制单元均值图，提供了检查加工过程影响的一种简单方法。替代的方法是在拟合包含测量轮影响的模型之后，绘制基于单元顺序编号的残差图。由于单元均值可能受到测量轮差异的影响，这在平衡嵌套设计后可能特别有用。应用数学校正来消除加工标准样品所产生的趋势是不合适的。如果检测到这种趋势并且发现该材料仍然是有用的，则均匀性不确定度宜以这种方式进行评估，即既要涵盖观测到的趋势、也要涵盖围绕该趋势的单元间任何随机变异。

7.7.3 单元间项的评定——基本设计

7.6.2 中描述的基本设计可以通过单因子方差分析进行分析，如附录 B.1 所述。计算假设有 a 个 RM 单元、对 RM 单元 i 进行 n_i 次测量，其中 $i=1, 2, \dots, a$ 。通常，简单的均匀性研究是所有 RM 单元具有相同重复测量次数 n_0 的平衡设计，因此， $n_1 = n_2 = \dots = n_a = n_0$ ，如果使用按照附录 B.1 所述的单因子方差分析方法，单元间标准差 s_{bb} 则可以由组间均方 M_{between} 、组内均方 M_{within} 和每个单元重复测量次数 n_0 按公式(3)(采用附录 B 中的符号)计算：

$$s_{\text{bb}}^2 = \max \left[\frac{M_{\text{between}} - M_{\text{within}}}{n_0}, 0 \right] \quad \dots\dots\dots (3)$$

在这种简单的情况下，单元间方差 s_{bb}^2 等同于单元间均匀性引起的不确定度(平方) u_{bb}^2 (见 7.11^[16])。

注 1：附录 C 中示例 C.1 描述了使用单因子方差分析的均匀性研究的实例。

注 2：通常假设真实的 RM 单元的均值是正态分布，尽管在用公式(3)计算时不需要这个假设。同样，通常也可以假设单元内随机测量误差服从正态分布^[13]。

注 3：如果不符合 7.5.1 中公式(2)所述的条件，请参见 7.8。

7.7.4 单元间项的评定——随机区组设计

某个随机区组设计，包含 a 个 RM 单元，每个单元有 n_0 个试样，且每轮每个试样仅重复测量一次共测量 n_r 轮并进行 n_r 测量，可以使用没有重复测量的双因子方差分析方法进行分析。统计模型为：

$$x_{ij} = \mu + A_i + B_j + \epsilon_{ij} \quad \dots\dots\dots (4)$$

式中， μ 是总体平均值， A_i 是单元 i 的效应， B_j 是轮 j 的效应， ϵ_{ij} 是随机误差项。除了增加了轮 j 的效应 B_j 之外，这与基本设计(附录 B.1)的模型相同。如果精密度依赖于观测的均值水平，尽管较大的轮次偏倚理论上可能会间接影响随机误差，但单元间方差 s_{bb}^2 的测定不依赖于 A 或 B 的分布。然而，通常轮影响大到足以表明这种影响会使整个研究完全无效。此外，分析假设轮影响和单元影响是独立的和叠加的。对于均匀性研究中预期相对小的影响(与组内和组间标准差相比)，这也是一个合理的近似值。

应用于随机区组设计的方差分析,按照公式(4)中的模型,每轮中每个单元观测一次,与之前一样,可得到轮间均方 M_B 、单元间均方 M_A 和残差均方 M_{within} 。

方差分析表也包括每项的自由度。ANOVA 表中自由度为 $(a-1)(n_0-1)$,而不是基本设计中的 $a(n_0-1)$,但残差均方 M_{within} 是重复性方差 s_r^2 的无偏估计。单元间标准差 s_{bb} 按公式(5)计算,自由度为 $(a-1)$ 。

$$s_{bb}^2 = \max \left(\frac{M_A - M_{\text{within}}}{n_0} \right) \dots\dots\dots (5)$$

注:随机区组设计处理的示例可参见附录 C 中示例 C.3。

7.7.5 单元间项的评定——平衡嵌套设计

平衡嵌套设计²⁾将 a 个单元平均分布在 n_r 轮中,每轮有 m_w 个单元,其单元间方差可使用 ISO 5725-3^[14]中描述的包括嵌套设计的方差分析方法进行计算,或遵循其他标准统计方法(如参考文献[15])。统计模型由公式(6)给出:

$$x_{ijk} = \mu + A_{ij} + B_j + \epsilon_{ijk} \dots\dots\dots (6)$$

式中, x_{ijk} 是轮 j ($j=1,2,\dots,n_r$) 中单元 i ($i=1,2,\dots,m_w$) 的第 k 次观测值, A_{ij} 是单元 i 在轮 j 中的效应, B_j 是轮 j 中的效应, ϵ_{ijk} ($k=1,\dots,n_0$) 是通常的随机误差项。平衡嵌套设计的方差分析同样会得到轮间均方 M_B 、单元间均方 M_A 和残差均方 M_w 。假设嵌套设计到每单元进行 n_0 次重复测量,则使用公式(3)再次估计单元间项,但具有 $a-n_r$ 个自由度而不是基本设计中 $a-1$ 个自由度。

附录 B.2 中描述了嵌套设计的双因子方差分析。将 a 个单元平均分布在 n_r 轮中的 a 个单元且每个 RM 单元进行 m_w 次观测, a 、 n_r 和 m_w 分别替换了附录 B.2 中的 p 、所有 b_i 和所有 n_{ij} 。

7.7.6 其他均匀性设计和可替代的估计方法

其他设计通常需要使用能够对各种设计进行方差分量分析的专门的统计软件,此类软件根据计算的均方并将该均方等同于预期均方继续用于方差估计;这是在 7.7.3 至 7.7.5 中使用的方法。然而其他统计方法已经可以获得,其中最感兴趣的是极大似然法^[15]和贝叶斯分析^[17]。虽然详细的计算超出了本文件的范围,但下面给出了选择此类方法的一些指导。

极大似然法是最普遍和广泛使用的估计方差分量的方法。标准样品均匀性研究中的方差估计推荐使用限制极大似然估计(REML)。这不仅提供了与无缺失数据的平衡设计方差分析基本相同的结果,而且很好地处理了复杂设计、缺失数据和明显不平衡。因此,REML 可用于代替上述基于方差分析的方法,也可用于更广泛的情况。

随着高级计算方法越来越可获得,贝叶斯分析越来越受欢迎。然而,贝叶斯分析依赖于对被估计参数(包括方差参数)先验概率分布的使用。它还通常需要仔细确定统计模型,并且一般来说计算量更大。虽然贝叶斯方法在经验丰富的人手中给出了良好的结果,但是当模型被错误地确定或选择不合适的先验概率时³⁾,特别是对于标准样品生产中常用的较小的数据集,也可以给出非预期的和误导的结果。因此,用于标准样品定值中方差分量估计的贝叶斯方法,宜仅在贝叶斯较小数据集分析方面经验丰富的合格统计学家的帮助下使用。

7.8 测量程序重复性欠佳的情况

由于测量程序的重复性要求(7.5)并不能总是得到满足,在这些情况下,可以使用另外一种替代方

2) balanced nested designs 也可译为完全嵌套设计。

3) 事件还没有发生,根据以往的经验给出的事件发生的概率。例如,根据以往经验样本单元的单元内标准差是 0.01。

法来估计单元间方差的上限。

参考文献[16]和[78]对测量程序重复性欠佳时获得不确定度估计的各种方法进行了讨论。参考文献[16]推荐的应用示例见附录 C 的示例 C.2。

7.9 单元内均匀性

7.9.1 评估单元内均匀性研究的需求

当 RM 使用说明书中规定的最小取样量小于整个 RM 单元量时将产生单元内均匀性的问题。确保充足的取样量是很重要的,但仅单元间均匀性研究并不能提供这样的保证。如果最小取样量较小,并且单元内存在明显不均匀性的高风险,则宜进行单元内均匀性试验研究。涉及明显单元内不均匀性的高风险情形包括以下一个或多个:

- 最小取样量明显小于单元的量;
- 通过混合粉末或颗粒材料制备的材料;
- 通过将少量的一种组分混合到大量基体中制备的材料;或
- 具有之前已知的单元内不均匀性的材料。

试验评估可能包括单元内不均匀性显著性检验或测定最小取样量。

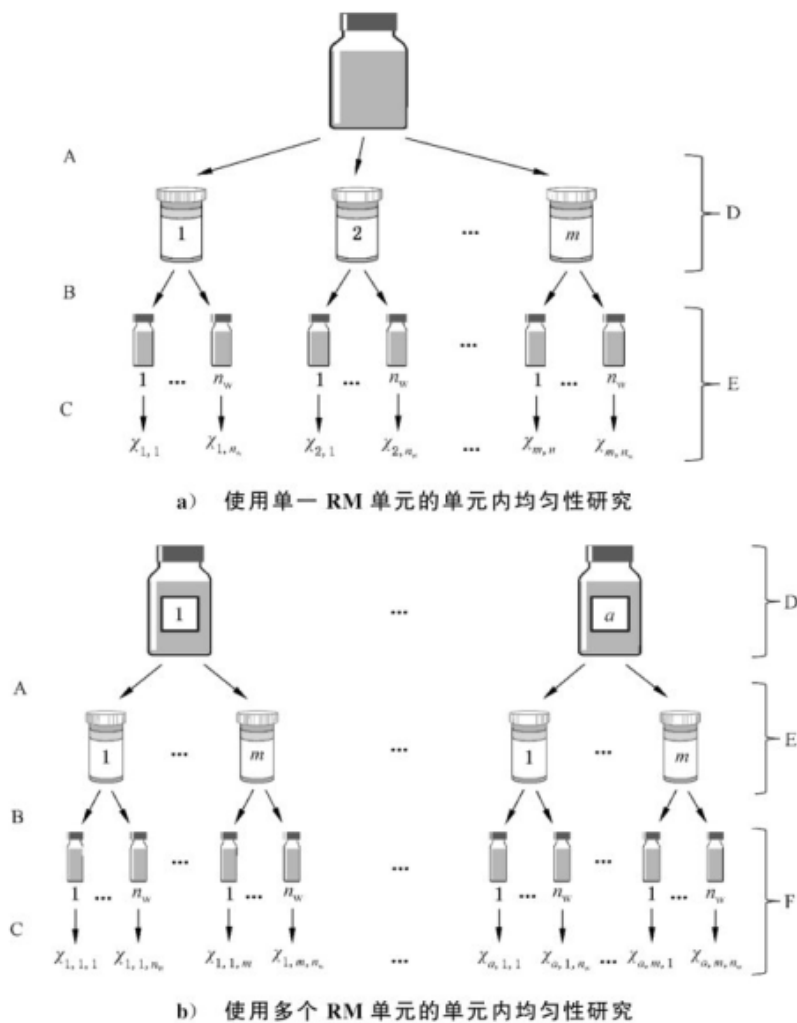
7.9.2 单元内不均匀性显著性检验

对于单元内不均匀性显著性检验,使用与单元间均匀性研究相似的设计,所不同的是有关方差是单元内子样项间的方差。图 5 是单元内均匀性研究的一些可能的设计。最简单的设计[图 5a)]是,从单个 RM 单元中依照所设想的最小取样量范围取出 m 份检验用样,并且每份检验用样被制备和测量 $n_w \geq 1$ 次。如果从单个 RM 单元不能取出足够多的试样,或者为了获得更多代表性的信息,通常随着单元内重复测量的减少,可包括更多的 RM 单元[图 5b)]。

由于单元内不均匀性检验宜为单元内项提供至少 5 个自由度。可以这样进行试验安排:例如,通过从单个单元[图 5a)中, $m \geq 6$]取 6 个或更多个子样;从 3 个或更多个单元中每个单元分别取 3 个子样[图 5b 中, $a \geq 3, m = 3$];或从 5 个或更多个单元中每个单元取 2 个重复子样[图 5b)中, $a \geq 5, m = 2$]。如果子样取自多个单元,则分析宜以类似于 7.7.5 平衡嵌套设计中包含轮项的方式包括单元间项。

如果可以对每个子样进行重复分析,则可应用方差分析,然后对子样间方差的显著性进行 F 检验。从单个 RM 单元中取出的各试样进行多次测量,可以使用如附录 B.1 中带有子样的单因子方差分析作为分组因子;如果从多个单元取子样并取出多份试样[图 5b)],双因子嵌套设计方差分析则是合适的(附录 B.2)。如果每个子样只能进行一次观测($n_w = 1$),可将观测值的方差(或来自多个单元的单因子方差分析组内项)与测量程序重复性的可靠估计进行比较,通常使用 F 检验找出不同子样中方差较大的结果。对于具有高自由度且可重复性方差估计良好时,可以将观测的方差与已知的重复性方差进行比较,使用卡方检验 χ^2 进行评估。

注:即使可以在子样上进行重复测量,也不一定能够进行如图 5 b)所示的子样的重复制备。



标引符号说明：
A —— 取子样；
B —— 制备；
C —— 测量；
D —— 用观测到单元间变异；
E —— 用观测到单元内、子样间变异的操作；
F —— 用观测到子样内变异的操作。

图 5 单元内均匀性研究设计图

7.9.3 评估最小取样量

7.9.3.1 试验测定

最小取样量可通过对不同大小取样量的单元内均匀性研究试验确定。由于单元内均匀性标准差取决于载有某种特性粒子的数量，因此，观测到的标准差预计会随着样品取样量的减小而增加，并随着样品取样量的增加而趋于测量程序的重复性。这可构成样品取样量研究的基础，步骤如下：

- a) 根据不同样品取样量(子样的质量或体积)下，计算由重复测量值得到的标准差的变化范围；
- b) 绘制标准差与样品取样量的散点图；
- c) 任选一个：
——使用图来证明样品取样量对评估范围内观测到的标准差没有影响；或

——选择某个样品取样量水平,该取样量得到的标准差与测量程序的重复性标准差无差别。

由于自由度较小,如果有明显的单元内不均匀性,进行试验的样品取样量范围要足以产生显著性差异。因此,如果可能,在此类试验中,样品取样量宜至少涵盖五倍的范围(如,从 10 mg 到 50 mg)。

如果可以证明子样间方差与取样量成反比,则英格梅尔(Ingamell)的抽样理论适用^[17]。在这种情况下,获得所需的量 m_1 与特定的子样之间的相对标准差 s' 相关,由公式(7)给出:

$$m_1 = \frac{K_1}{(s')^2} \quad \dots\dots\dots (7)$$

其中, K_1 是英格梅尔抽样常数。

注:英格梅尔抽样常数是引起子样间 0.01(1%) 相对标准差的材料的质量,可以由其他子样量的子样间相对标准差来测定。

7.9.3.2 其他方法

最小样取样量可使用其他数据或经验进行设置,能够用来选择最小取样量的其他数据包括,例如:

- 在稳定性研究中用于获得具有可接受精密度结果的样品取样量;
- 在单个实验室进行的定值研究中提供的可接受精密度的样品取样量;
- 在实验室间定值中获得可接受精密度的实验室的样品取样量;
- 用于均匀性研究的样品取样量。

7.10 均匀性的确认

在确定了单元间和(适用时)单元内标准差之后,标准样品生产者宜确认单元内和单元间变异对材料的预期用途足够小。这种确认可以包括,例如:

- a) 将单元间和/或单元内标准差和与定值相关的不确定度进行比较,确认标准差与定值不确定度相比较小(如, $s_{bb} < u_{char}/3$);
- b) 在计算标准值的合成不确定度 u_{CRM} 时考虑了不均匀性带来的不确定度,并确认标准值的不确定度对预期用途是可接受的;
- c) 对于不计算合成不确定度 u_{CRM} 的标准样品,核查单元间标准差 s_{bb} 与使用领域中典型的实验室间再现性标准差 s_R 相比是否较小。理想情况下, s_{bb} 宜小于 $s_R/3$;
- d) 例如通过使用 F 检验法,确认在 95% 的置信水平时单元间项不具有统计显著性。

注 1:无统计显著性不能够证明不存在不均匀性,但是当它不是为了提供特性值的不确定度时,可能是一个有用的实用准则,并且用于均匀性研究的程序是标准样品最终用户使用的典型程序。

注 2:ISO 13528^[10] 中给出了单元间标准差不显著大于预定上限的方法。

7.11 均匀性研究的不确定度评定

标准值的不确定度(见 10.2)包含由不均匀性带来的不确定度分量 u_{hom} ,包含所有检测到的不均匀性来源。 u_{hom} 不宜小于由均匀性试验确定的单元间标准差 s_{bb} 。如果相比于其他不确定度分量, s_{bb} 可以为零或可忽略不计,此时可以忽略 u_{hom} ,除非单元内标准差更重要,否则可以不考虑 u_{hom} 。

如果在使用说明书中推荐的最小取样量时存在显著的单元内不均匀性,与标准值相关的不确定度则宜包含 u_{wb} ,且不小于试验研究中得到的子样间标准差(见 7.9 和关于 u_{wb} 估计的附录 B.2 注)。那么,与不均匀性相关的不确定度 u_{hom} 由公式(8)给出:

$$u_{hom} = \sqrt{u_{bb}^2 + u_{wb}^2} \quad \dots\dots\dots (8)$$

为了允许精密度不足的测量程序用于均匀性研究,可以增加 u_{hom} 对于标准值不确定度的贡献,如 7.8 中所述。

8 稳定性评估与监测

8.1 概述

关注特性的稳定性是所有 RM 的关键特征之一。重要的是,每个关注特性的值在使用时要与所附该 RM 的证书或其他文件中所陈述的值相一致。

每个特性的值可能因各种原因随时间而变化,其变化程度和变化速率取决于条件。三种类型的条件非常重要:在标准样品生产者设施中长期储存条件,运往用户场所期间的条件,以及在用户场所储存和使用的规定条件。

不同材料的变化形式和变化速率存在很大差异。某些在各种条件下变化很小或者根本没有变化,某些在常温下快速变化需要低温储存。降解的形式可能不同:某些材料在长时间内几乎呈线性变化;某些在稳定期之后,可能会进行自催化或难以预测的快速变化;某些在加工后的初始阶段迅速变化,然后保持长时间稳定。对某些材料来说,变化对材料的某些单元的影响比其他单元大得多,导致单元间均匀性的变化,而不是所有单元的值的一致变化。

这些在不同的条件下迥然不同的变化模式,即使经过长期的试验研究也很难预测。因此,在标准样品寿命中、或甚至在常规生产的某个特定批次标准样品寿命中,总有一个或多个特性的值存在发生意外变化的风险。所以,期望标准样品生产者管理材料加工、储存、包装、运输条件、证后监测和给最终用户的建议,使得意外变化的风险尽可能最小。

RM 生产者宜:

- a) 必要时通过试验评估标准样品在拟定储存条件下所有相关特性的稳定性,并据此选择预处理、包装和储存条件;
- b) 必要时通过试验评估标准样品在计划运输条件下所有相关特性的稳定性,并选择运输条件以确保运输期间的稳定性;
- c) 制定材料储存和使用的必要建议,以确保在用户场所的稳定性;
- d) 考虑到可能的变化率,选择一个能够及时检出变化的长期储存材料稳定性的监测方案;
- e) 如果不能确保标准值的稳定性,则宜考虑使用前标准值的可能变化,对不确定度进行适当的补偿,或者当随时间的变化可以预测时,提供标准值及其不确定度随时间变化的修正方法;
- f) 当使用说明允许从标准样品单元中重复取样,或者整个标准样品单元重复使用时,评估对材料稳定性的可能影响并采取适当的措施。

以下条款对试验研究的必要性,以及需要试验研究时对其内涵和实施提供了相应指导。此外,还对随着时间推移仍有明显不稳定变化风险的情况,给出了相关不确定度的必要性和评估的指导。

注 1: 符合以上 a) 到 f) 条款是 ISO 17034 的要求。

注 2: 预计对于大多数材料, RM 生产者将通过选择预处理、包装、储存、运输条件和监测方案的组合,达到随时间推移可忽略变化的合理预期。

注 3: 监测结果还可以为今后相同或者相关 RM 的稳定性管理提供参考。

注 4: 任何针对随时间推移预测变化值的方法,对于标准值,还需要评估修正值的不确定度。

注 5: 某种材料,在随时间测量时,即使其他特性值可能会有明显的变化,也可能有一种特性是稳定的。

注 6: 大多数描述数据随时间变化的统计处理方法都要使用数学模型来评估连续变化。然而,虽然可以通过给定研究的时间过程检查一个模型的适用性,但是未来可能的降解形式非常广泛,这妨碍了对降解的可靠预测。将不确定度分量纳入可能的替代降解模型中也是不实际的;不同模型之间的巨大差异使得这种不确定度变得不切实际的大。实际上,在给定的条件下,估计长期变化的唯一可靠方法是在整个关注期内对材料进行观测。因此,无论使用何种模型来检验研究过程中的变化和预测未来的变化,稳定性监测对于任何打算长期使用的材料来说都是至关重要的。

8.2 稳定性评估

8.2.1 稳定性评估的要求

所有 RM 的稳定性均宜进行评估。稳定性评估可能涉及试验研究,以估计加工后候选 RM 的不稳定性程度,或者用来确认该材料的稳定性。因长期不稳定性或运输的影响引起特性值可检出的任何变化,宜在宣称特性值时加以考虑,并且对这些变化估计的不确定度宜包含在标准值的不确定度中。

如果标准样品以重复批进行生产且各批没有单独进行稳定性检验时, RM 生产者还宜对从一批到下一批间稳定性变化的风险进行评估,并宜通过试验验证足够数量的不同批的稳定性,以对所有批的稳定性提供信心(见 8.2.4.2)。

8.2.2 (不)稳定性的类型

标准样品生产中,两种类型的(不)稳定性十分重要:

- 材料的长期稳定性(即在规定的储存条件、有效期内材料的稳定性);
- 合理预期运输条件下的稳定性(“运输稳定性”)。

标准样品的长期稳定性与其在 RM 生产者的库存或者使用者按照规定条件储存时 RM 的状态有关。在标准样品分发给用户之前,宜对材料的长期稳定性进行评估。

运输稳定性与样品运输带来的任何额外影响有关。在某些情况下,不可能在运输过程中始终保持 RM 稳定性的适当条件,此时,宜在标准值不确定度中给予一些适当补偿。在向用户分发标准样品之前, RM 生产者宜评估每个关注特性在合理预期运输条件下的稳定性。

注 1: 运输稳定性通常被称为“短期稳定性”。

注 2: 如果不能保持预期的运输条件,了解 RM 可能会发生什么情况通常是很有用的。该知识可使 RM 生产者能够提供更好的建议,而且从用户的角度来看,得到了一个更好的产品。因此,可以在比预期储存和运输条件更加极端的条件下评估不稳定性。

8.2.3 稳定性评估的一般方法

长期和运输两种稳定性评估宜包括下列一项或多项:

- 考虑材料中有可能影响稳定性的所有物理、化学和生物特性,包括经过认定的特定化学或生物物种;
- 查阅已发表的相关材料稳定性的文献资料;
- 查阅相关材料稳定性评估或稳定性监测数据;
- 策划试验,无论是实时的或加速的稳定性研究;
- 试验以检验不同储存安排的影响,包括容器的完整性、稳定化或保存方法。

当使用稳定化措施时, RM 生产者宜说明稳定化的方法,并验证(通过上述任一种方法)该稳定化方法适用于材料的使用且有效^[18]。

当在规定使用条件下允许重复使用或重复取样(如重复开瓶、重复冻结或者湿度的影响)时,稳定性评估宜考虑其潜在影响。如果不能评估重复取样或者重复使用情况下的稳定性,宜在使用说明中禁止重复取样或者重复使用。

注 1: 上述“相关材料”可以包括与候选标准样品明显不同的材料,这些材料具有某些(但不是全部)相关特征。例如,在不同基体或者溶剂中相同的定值特性,相同基体类型但不同的定值特性,等等。

注 2: 8.2.4 和 8.9 中包含了重复取样或重复使用的进一步说明。

8.2.4 稳定性试验研究的需求

8.2.4.1 稳定性试验研究的一般要求

RM 生产者如标准样品在计划储存和/或运输条件下稳定性的先验信息掌握很少或没有,或者对

重复取样或重复使用的影响不清楚时,则宜进行稳定性试验评估。

如果 RM 生产者在相同的计划储存条件下,从长期保存的非常相似材料中获得了稳定性的先验信息,则无需进行试验研究。

8.3 至 8.8 对稳定性试验研究的形式和实施提供了指导。

注:“非常相似”的材料是定值相同特性的材料,它们有相同的基体组成、制备条件、相似或更加有效的包装等。

8.2.4.2 重复批生产的稳定性试验研究需求

除了选择用于稳定性试验测定的批次外, RM 生产者在以下情况时,通常还宜通过试验验证至少两个批次的稳定性:

- 试验研究被认为是必要的;
- 有关材料是重复批次生产的;
- RM 生产者拟使用一个批次的稳定性数据代表后续批次的稳定性。

注:验证可以是一种简单的检验,以确认不同批次的稳定性状况相似,或者对于连续批次,在其寿命内不发生变化,而稳定性试验评估通常涉及一项旨在确定变化速率的扩展研究。

8.3 稳定性研究的分类

8.3.1 概述

稳定性研究的方法可能有很大的不同。最简单的方法是在不同的时间点进行两个独立测量,以便根据经过时间的变化得出关于未来稳定性的结论。其他方法包括多点方法,该方法使得 RM 生产者能够对更复杂的稳定性状况建立模型,并/或得出比稳定性研究中第一次至最后一次测量之间经过的时间更短期变化的结论。

稳定性研究可分为:

- 根据测量条件,分为经典和同步研究;
- 根据应用条件和处理方法,分为实时研究和加速研究;
- 根据是否要评定运输条件稳定性还是长期储存稳定性,分为运输和长期稳定性研究。

此外,还需要检查特定的条件,如包装特点或湿度,这可能需要特殊的试验设计。

以下各条是对这些不同设计选项的指导。

8.3.2 根据测量条件分类

8.3.2.1 经典稳定性研究——中间测量条件

经典稳定性研究中,在相同条件下同时制备的各样品(即作为一批样品),按时间推移进行测量(例如,一个样品立即测量、一个在 3 个月之后测量、一个在 6 个月之后测量,等)。当测量系统的不稳定性对测量结果的离散性影响很大时,这种在中间测量条件下(有时被称为实验室内再现性条件)进行测量的设计,会导致相对较高的不确定度。

8.3.2.2 同步稳定性研究——重复性测量条件

同步设计使用参考条件储存,以允许暴露在不同的降解条件下的 RM 单元在很短时间内、在理想的可重复条件下进行测量^[19]。参考条件是一组条件,在该条件下能可靠地预期有关特性是稳定的、或在选定的基线水平。“同步”一词强调测量是在同一时间进行,而不是像经典稳定性研究方法那样,分布在稳定性研究的整个时间跨度上。这种重复性条件的使用,预期将改善试验过程中的测量精密度,从而改善稳定性研究的功效。因此,理论上,同步稳定性研究比经典研究的不确定度更小,这取决于测量的重复性和(实验室内)再现性之间的差异。该设计的前提是,在给定条件下材料不发生降解,或者至少与

选择的储存条件的降解速率有差别。

8.5.3 中对同步研究进行了更详细的描述。

8.3.3 根据稳定性研究持续时间和条件分类

8.3.3.1 实时稳定性研究

在实时稳定性研究中,某种材料的稳定性是在 RM 预期的储存或运输条件下进行研究的。这意味着一周/月/年稳定性研究提供了该材料在过去一周/月/年的稳定性信息。这类研究的优点是它不需要任何关于不同条件对稳定性影响的假设,因为用于试验研究的条件与预期用于运输或储存的条件是一样的。

8.3.3.2 加速稳定性研究

在加速研究中,在比 RM 预期储存或运输条件更加极端的条件下进行多个试验,目的是诱导比预期储存条件更快速的降解。然后,通过诸如对各种试验条件的回归分析,估计关注条件下材料的降解速率。最常见的例子是在几个温度下进行检验,并通过阿伦尼乌斯(Arrhenius)方程估算降解速率。这类研究的优点是缩短了所需总时间,并因使用更极端条件对结果增强了信心,主要缺点是降解机理或速率决定步骤会随研究条件的不同而改变,尤其是温度。

注 1: 实时研究和加速研究都可以在经典或同步研究中使用。

注 2: 使用加速研究可以提供比研究持续时间长很多时间内的稳定性信心,并且在需要提前获得该材料时特别有用。

注 3: 光、水分和温度是能加速材料降解的常见因素。

8.3.4 根据研究目的分类

8.3.4.1 运输或其他短期稳定性研究

运输稳定性是材料的一种特性,是指在预期运输条件下的稳定性。为此,需研究在(至少)预期包装和运输条件下材料的状况及其特性值。运输的预期条件越严格,运输稳定性研究中需要包括的条件范围越小。

当没有关于某种特定类型材料的之前经验时,宜进行运输稳定性研究,以获取合适的运输条件信息。运输稳定性研究中包含的持续时间和条件宜能够反映将 RM 单元运输到使用者场所合理预期的持续时间和条件。除非有相反的理由,这些条件宜包括在国际运输期间可能合理发生的极端温度,并且持续周期至少是运输 RM 允许的时间。例如,如果建议运输时间限制在 3 周内,短期稳定研究 3 到 4 周就足够。

如果有类似材料稳定性的经验,并表明需要严格控制运输条件(如用于冰或冰袋包装的运输),研究可能仅限于预期的运输条件。

其他短期研究可能也是必要的,例如建立最终用户场所预期储存条件的适用性。在最终用户场所确定短期替代储存适用性的研究,宜反映使用说明中允许的条件(见 8.9)。为了达到这个目的,通常不需要考察更严格的条件。

注: 在本文件中,假设运输条件通常都会得到充分严格的控制,并且运输的影响不会显著影响特性值,也不会增加标准值的不确定度。因此,假设运输稳定研究的主要目的是核查显著变化,而不是对变化提供准确的定量估计。

8.3.4.2 长期稳定性研究

开展长期稳定性研究,是为了评估在产品寿命规定的储存条件下的稳定性。实时长期研究通常持

续 12 个月或更长;加速研究通常较短,但包括极端条件。证书发布之后的有效期仍然需要通过稳定性监测来保证。一般来说,某个材料中特性值可获得的稳定性数据越少,越需要加强长期稳定性和证后监测。当新 RM 急需上市时,可以将长期稳定性研究限制在 12 个月以内,并增加证后监测频次,以弥补认定之前数据的不足。

8.3.5 不同储存和处理条件的设计

有时有必要检查一种以上储存或处理条件的影响。例如,可能需要评估高湿度、高低温度、不同的稳定化方法、光照的影响,或不同的包装选项。

为此目的,试验可能包括:

- 对稳定性每种可能影响的单独研究;
- 多因子试验,旨在同时评估几种储存和/或包装处理的效果;
- “最糟糕情形”研究,将材料暴露在最极端的条件组合下,根据最极端条件下的稳定性提供在不太极端条件下的稳定性证据。

如果需要研究多种储存和/或包装条件,出于效率的原因,鼓励使用析因设计或部分析因设计。使用时,这类试验宜考虑不同因子之间相互作用的可能性。两因子的相互作用,如湿度和温度的组合效应,通常是很重要的,试验设计宜允许在可行的情况下对其进行估计或检测。

注 1: 析因和部分析因试验设计的指导见参考文献[11],其在(药物和产品)稳定性研究中的应用描述见参考文献[20]。

注 2: 如果在多因子试验中发现某种影响,通过进一步的研究验证该影响是重要的。

注 3: 在试验设计文献中,允许从双因子交互作用中无偏估计主效应的部分析因设计有时称为“分辨率 IV”设计。分辨率 V 设计还允许某个双因子的交互作用与其他双因子交互作用分离。

8.4 有效稳定性研究的一般要求

8.4.1 要求概述

为了在稳定性研究中获得可靠的结果,重要的是:

- 选取具有代表性的样品单元组;
- 选择一个具有足够精密度和选择性的合适测量程序;
- 按照适当的试验设计在适当的条件下进行测量;
- 使用有效的统计方法进行统计分析。

8.4.2 至 8.4.4 为上述前三点提供了指导,8.5 详细讨论了统计分析。

8.4.2 单元的选择

用于稳定性研究的制备和包装材料单元通常宜在包装单元中随机选取。

但是,可以用较小批的材料进行初步稳定性研究,以便确定适当的包装或预处理。在这种情况下,专门制备用于初步研究的单元并研究所有制备的这些单元或单元组就足够。

注 1: 相比于均匀性研究或赋值,稳定性研究的解释通常更少依赖于对单元的随机选取;关于稳定性有用的指示性信息很可能从简单的抽样策略(如系统抽样)中得到。然而,重要的是要避免任何可能导致不具有代表性的稳定性信息的抽样方案。例如,从大批 RM 单元的外部区域选择单元,如大容器或储存区域的顶部或侧面,导致选择了暴露于光照或更极端温度变化中的样品。这些单元很可能已经经历了一些变化,并在进一步的稳定性评估中表现得不具有代表性,特别是当该材料容易产生早期快速变化并在之后近似稳定时。

注 2: 7.4.3 给出了选取 RM 单元用于研究的适当抽样方案的示例。

8.4.3 用于稳定性研究的合适测量程序

无论研究设计如何,只有在研究时间跨度内的测量结果的标准差(可能与单元间均匀性有关)足够

小时,结果才有意义。

对于主要依赖于良好重复性的同步研究(见 8.3.2.2 和 8.5.3)或其他设计所使用的测量程序宜主要选择具有良好重复性的。对于可能受到测量中轮-轮变异影响的研究,例如,单一储存条件下的简单经典设计,宜首选具有良好中间精密度的测量程序。

8.4.4 合适的试验设计

8.3 给出了稳定性试验研究的一系列设计。如果可以证明其有效性,替代设计也可以使用。

稳定性研究中要求重复的程度和研究的持续时间取决于多种因素。特别是,预测降解的不确定度既取决于研究持续时间,也取决于重复的次数。在确定重复次数、不同暴露时间的数量、不同组暴露条件的数量和范围时,宜考虑以下方面。

- 被选择用于研究的测量程序能达到的精密度;非常精密的测量程序可要求较低的重复性水平;
- 允许个别观测或 RM 单元出现失败的需求。对于时间和温度(或其他条件)的每种组合,允许至少有两个 RM 单元以提供冗余量和有助于避免遗漏条件。
- 材料的计划寿命和监测频率。长的监测间隔需要时间长的稳定性研究或更高的重复性,以提供可靠的降解预测。
- 验证线性(或其他)模型效果的需求。为了对线性度进行检查,在同步研究中至少需要进行三次观测;对于非同步研究,如果轮影响可能是重要的,至少需要在四个时间点进行观测,而且要求重复测试。
- 在建议的储存条件下观测材料的需求。研究宜包括(但不限于)暴露在建议的储存条件下。
- 研究影响稳定性的因素之间相互作用的需求。

除上述考虑外,对于之前可用信息较少的材料,所使用的单元数和重复次数宜足以为预测到的降解提供较小的不确定度。

如果可用的测量程序没有足够的精密度来进行可靠的稳定性测定,则在每个时间点和条件组合下宜包含额外重复次数和/或额外的时间点。重复的首选性质取决于变异的主要来源,如下所示。

- 当测量重复性是变异的主要来源时,宜增加在时间和其他条件组合下每个单元的重复测量次数和/或研究的单元数。
- 当 RM 不均匀性(以单元间标准差 s_{ub} 表示)是一个重要的变异来源时,宜增加每个时间/条件组合下研究的单元数。
- 在经典稳定性研究中,当随时间的测量变异很重要时,宜增加时间点的数量。

注 1: 当单元间不均匀性由已知趋势引起时,将 RM 单元分配到不同时间/温度进行处理以避免偏差很重要,这通常通过确保随机分配单元进行处理来实现。

注 2: 在稳定性研究中,暴露时间不一定要等距分布。

注 3: 国际协调会议(ICH)描述了一些有效的设计,用于研究影响稳定性的多种因素^[20]。

8.5 评定稳定性研究结果

8.5.1 对稳定性研究数据处理的一般考虑

稳定性研究的数据处理宜考虑到特定的研究目标、所使用的试验设计以及可能影响结果的变异来源。

对于大多数基本稳定性研究,目标既是检测在储存过程中随时间的任何重要变化,也是估计特性值随时间的变化速率。在研究单个或多个储存条件(上述 8.3.5)的影响时,数据分析通常为显著影响提供检验。一些设计中包含了这两个目标的元素。

所使用的试验设计将影响数据分析选项。仅限于两组条件(如两个时间极点)的设计既可以利用两组(假设每个点进行重复测试)之间显著性变化的检验、也可以通过回归分析来进行评估,而在这种简单

情况下,这两个程序是等效的。跟踪特性值在一组储存条件下随时间变化的设计,通常使用建模(如线性回归,8.5.2.3)来评估,以估计变化的速率或通过方差分析来检验随时间推移的显著性差异,但是无法通过简单的两组检测进行评估。跟踪某个材料在几种储存条件下随时间变化的加速研究,可能需要使用一个更通用的模型来进行处理,该模型可以估计在不同条件下的不同变化速率(如降解速率对温度的关系)。

影响结果的随机变异也影响数据分析。对于最简单的设计,其中只有一个随机变异源,8.5.2 描述的简单线性回归通常就足够。对于涉及不同时间点测量的研究,其中除了轮内变异之外,测量系统还可能出现轮间变异,选择的数据分析方法宜可以处理存在其他来源随机变异的情形。此类情况的数据分析在 8.5.5 中有所考虑。

以下各条描述了稳定性研究数据的分析。8.5.2 描述了线性回归在最简单情况下的应用,并提供了对更复杂情况的指导,包括使用一种储存条件和一种参考条件的同步研究,加速研究和具有更复杂误差结构的研究。

注:有证标准样品不确定度评定也能够使用稳定性研究的结果。

8.5.2 经典稳定性研究:单一储存条件下的多个时间点

8.5.2.1 适用性

本条款适用于在多个时间点中每个点进行一次或多次测量的情况,每次测量的随机误差是独立的,并且具有相同的标准差和分布(即误差是独立的并且分布相同)。

注 1:当在不同时间测量且测量系统的中间精密度标准差大于重复性标准差时,独立性假设不成立。更多信息见 8.5.2.5。

注 2:如果从一个时间点到另一个时间点的精密度发生显著性变化时,将不满足服从相同分布的假设,本条款不适用。精密度的变化可以通过例如均匀性方差或残差图来检验。更多信息见 8.5.2.5。

8.5.2.2 程序

对于基本情形,使用简单线性回归,程序如下:

——拟合初步模型,检查拟合模型残差,核查假设(详见 8.5.2.4 和 8.5.2.5);

——如果假设适用,记录模型参数(通常是斜率和截距)和不确定度,并核查发现的任何趋势的统计显著性(见 8.5.2.6)。

简单线性回归的详细计算方法已在附录 B.3 中给出,并且通常可以在所有通用的统计软件包和大多数商业电子表格应用程序中获得。

注:基本上所有的线性回归软件,包括商业电子表格应用程序,都会自动执行这些计算。ISO 17034 要求软件(包括 RM 生产者为此目的编写的任何软件)都要在使用前进行确认。

8.5.2.3 模型选择

对于潜在的动力学机理未知并且预期变化很小的稳定性评估,线性近似通常是合适的模型。附录 B.3 中描述了简单线性关系。

在明确定义的非线性机理是不稳定性原因的情况下,相应的降解模型要优于(经验)线性模型。对于直线以外的模型而言,数学运算有些复杂,但评定运行方式相同。

示例:含有放射性同位素的 RM 是特性具有明确定义动力学机理的实例,在这种情况下的放射性衰变是可以通过公认的非线性模型来预测的。

8.5.2.4 拟合模型

回归参数可以使用附录 B.3 中的程序或合适的软件进行计算。计算分别提供了真实截距 β_0 和斜

率 β_1 的估计值 b_0 和 b_1 , 以及相应的标准误差 $s(b_0)$ 和 $s(b_1)$, 这些参数可用于后续的统计检验和不确定度评定。

8.5.2.5 检验和假设的核查

宜对回归结果进行检验, 并按照附录 B.3.3 所示对假设进行核查。

在标准样品稳定性研究中, 核查不合适的降解模型的证据特别重要。在简单线性模型中, 检查非线性的证据也特别重要。

如果发现严重偏离假设的情况, 宜终止数据分析, 解决异常情况或采用其他数据处理方式。一些常见问题和解决方法如下。

- 宜检查离群数据点, 并在适当情况下剔除。离群数据点对模型拟合和显著性检验均有不利影响。如果是测量系统故障, 离群数据点宜被校正或剔除。但是, 离群值也可能是由于选择了错误的降解模型, 或者是由于 RM 单元单独降解; 因此, 在剔除任何观测值之前, 宜谨慎考虑可能的原因。在稳定性研究中, 如果材料已经证明是均匀的, 同一单元在重复观测中出现单个离群观测值则可能表明测量不成功。
- 如果单元间存在显著的不均匀性(与可获得的测量精密度相比), 则有必要在统计评估中考虑到这一点; 特别是, 如果未考虑单元间不均匀性, 显著性则可能会被夸大。
- 测量系统中轮间变异的证据宜通过替代的数据处理方式解决(见 8.5.5)。
- 曲性模型的证据可以通过使用不同的模型来解决, 或者(因为可见的变化意味着不稳定性)用来断定材料对预期用途不够稳定。

8.5.2.6 检验显著不稳定性

对于上述简单情形, 通常稳定性检验是用 t 检验来检验斜率是否显著不同于零, 如附录 B.3.4 所述。

假如趋势具有统计显著性, 则需要考虑在技术上是否具有显著性; 也就是说, 是否需要增加标准值的不确定度或阻止认定。假如与讨论的特性值标准不确定度相比, 在标准样品有效期内预测的降解更加重要, 则可认为趋势在技术上是显著的。

如果观测到有技术上显著性趋势, 则适用 8.6 的规定。

8.5.3 同步设计

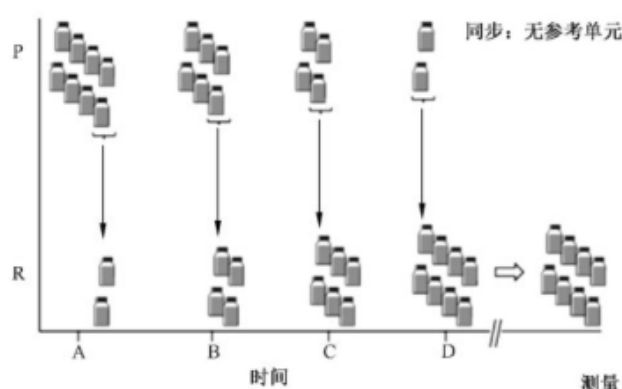
8.5.3.1 具有一个储存条件和一个参考条件的简单同步研究

在同步设计中, RM 的一组单元在建议的储存条件下暴露一段时间, 然后在达到每个单元的计划暴露时间之后, 分别转移到参考条件下。当所有单元达到预期暴露时间后, 测量整组单元。该顺序如图 6 所示。

通常, 计划储存条件为环境温度或接近环境温度(如 4 °C), 参考条件包括一个低很多的温度, 在该温度下认为降解基本不会发生。

由于研究仅涉及单一的暴露条件, 所以数据分析遵循 8.5.2 基本稳定性研究的程序, 使用每个单元暴露于计划储存条件的时间作为分析的时间轴。

注: 在同步研究中, 也可以将材料从参考条件移动到计划暴露条件, 以获得所需的暴露时间。



标引符号说明:

P——计划储存条件;

R——参考条件。

图6 同步研究示意图

注: 该图为同步研究示意图, 该研究包括在预定的研究期间等距增加暴露时间 A、B、C 和 D 之后进行观测。许多单元(这里有 8 个)被留作研究。在时间点 A(可能是名义上的零暴露时间), 将一组单元移动到参考条件。在随后的每个时间点将相等数量的单元移动到参考条件。最后, 在所有单元都完成暴露所需时间之后, 所有单元储存在参考条件中, 然后同时取出并测量。注意, 如果在短时间内完成测量, 则无需在参考条件下进行。

8.5.3.2 同步稳定性研究的优缺点

同步研究, 包括 8.5.3.1 的简单设计, 具有以下优点。

- 所有测量均在短时间内进行, 通常是在单轮重复性条件下进行。这为研究提供了最佳的可用精密度, 并避免了将测量系统的长期漂移误认为不稳定性的可能性。
- 在单轮中完成所有测量, 降低了成本并简化了实验室资源的调度。

缺点包括以下几点。

- 将测量推迟到研究结束, 意味着在研究完成之前无法确定不稳定性。
- 8.5.3.1 的基本设计中, 对结果的解释依赖于在参考条件下样品的变化微小。如果材料在参考条件下(例如, 由于逐渐凝结的影响)逐渐变化, 这既可能被误认为是在计划储存条件下的变化, 也可能会导致得出该材料是稳定的错误结论。因此可能需要额外的证据来进行可靠的说明。
- 该研究设计假设在材料进入或移出参考条件时没有不利的影响。如果材料在转换条件时发生变化, 结果则可能是高度可变或难以解释。
- 并非所有材料都可以放置在比预期储存条件更能有效防止降解的条件下。有些不能冷却到低温, 其他的可能需要在可达到的最低温度下储存。
- 不是所有的降解机理都会受到条件变化(包括温度)的影响。

8.5.3.1 中的基本设计还有一个附带的潜在缺点, 即仅包含了一个候选储存条件。

注: 如 8.5.3.3 所述, 可通过扩展设计来降低上述缺点。

8.5.3.3 基本同步设计的扩展

8.5.3.2 中列出的一些缺点可以通过扩展基本同步设计来解决。

- a) 如果需要, 运行不同持续时间的两组研究可以较早地提供结果。
- b) 包含多种替代暴露条件, 包括“加速”条件。这可以允许在一系列条件下评估稳定性, 还能允许对加速稳定性研究进行同步测量。此外, 包括在中间条件(即参考和计划储存条件之间)下额

外保留的 RM 单元可提供一些保护,以防止对参考条件下意外变化的误解。

- c) 使用随机区组设计进行测量。如果在单个测量轮中有太多观测要完成,则可以使用随机区组设计进行测量。在这种情况下,在几个测量轮的每轮中,对每种暴露处理影响以较小的次数进行观测。这以牺牲自由度的微小损失为代价,提供了根据重复性标准差进行稳定性相关变化的一种检验,类似于均匀性研究中使用随机区组设计(见 7.6.3)。

上述 b)和 c)涉及的扩展设计的数据分析比简单线性回归更复杂。对于多温度同步研究的处理,适用 8.5.4 的规定。对于随机区组设计,所使用的程序与 7.6.3 类似,通常需要统计软件;然而,对跨测量轮的每次处理观测值的简单平均值允许进行如上所述的简化评估。

8.5.4 多重暴露条件的加速稳定性研究

8.5.4.1 加速研究的概述

这类研究通过将受试对象放置到比计划储存条件更加极端的条件中,通过可能的影响因素加速降解效应的发生,从而缩短观测变化发生的时间^[21-22]和/或增加观测到的降解。

可能加速降解的因素包括但不限于:

- 环境因素,如温度、湿度、辐射(通常是光,尤其是紫外线)、氧化剂(如空气)、储存容器性质(接触面受污染、接触面吸附特性、渗漏);
- 样品特定的因素,如活性基体成分,特别是天然基体 RM。

由于温度在大多数已知降解过程中起主要作用,包括化学反应、扩散、蒸发、或吸附,加速研究通常宜对温度因素进行检验。研究的温度范围宜至少包括在计划的储存和运输条件下材料可能遇到的合理温度范围。

除温度以外的其他因素,特别是湿度和辐射,如果没有现有证据表明可以忽略,宜在加速稳定性研究中予以考虑。

尽管一般方法也可用于其他任何连续因素,但以下各条仅限于温度因素。

注:如果在加速研究中没有观测到任何变化,则不可能拟合出可靠的预测模型。因此,如果可以选择加速研究中包含的条件,至少能够在极端条件下观测到明显的变化,这对随后的数学建模和变化预测都是非常有用的。

对于医药产品,温度、湿度和辐射(在某些特殊情况下)能对材料的稳定性产生影响。国际协调会议(ICH)建议描述加速降解研究,其中受试材料将暴露在这些因素的某些确定组合之下^[20]。

示例:表 1 给出了某个候选材料加速老化研究的可能设计示例。根据预期的状况和降解速率,设计可能会包含更多或更少的时间和温度设置点,范围可能根据研究的被测量种类而有所不同。

表 1 加速老化示例:温度和暴露时间

温度/℃	4	20	40	70
暴露时间/月	1,3,6,12,(24)	1,3,6,12,(24)	1,3,6	0,5,1
注: 括号中的暴露时间显示了第一次证后监测的采样时间(见 8.10)。				

8.5.4.2 评估加速稳定性研究的数学模型

所讨论的分析物由温度驱动降解的时间依赖性 $g(t)$,通常是利用分析物随时间的变化率(dx/dt)与样品中可用分析物总量之间比例的基本假设进行建模,比例系数是与温度相关的反应速率 $k_{\text{eff}}(T)$ 。因此,从公式(9),得到与时间的指数相关式:

$$\frac{dx}{dt} = -k_{\text{eff}}(T)x \quad \dots\dots\dots (9)$$

x 是所讨论的被测量的值。尽管这种假设并不普遍适用,但它描述了大量的降解过程。因此,依据

这个假设建立的简单线性模型宜是 $\ln[x(t)]$ 而不是 $x(t)$ 本身,并且如下所述的回归宜在对数线性空间中进行,或者宜使用适合于所选模型的非线性曲线拟合。

在给定温度下,一个有效的伪一级反应速率 $k_{\text{eff}}(T)$ 可以由一个以上同时降解的过程产生,每一个都与不同的温度相关。温度驱动的降解过程通常遵循阿伦尼乌斯(Arrhenius)模型,生物/微生物过程可能遵循奥尼尔(O'Neill)模型,而样品中扩散驱动的变化可能遵循菲克(Fick)定律。如果有可能的降解过程的先验信息可用,动态反应动力学可以作为描述反应速率对试验因素水平的相关关系的基础。在大量单个过程重叠并且可以(至少部分地)补偿的情况下,经验模型(没有物理化学基础)是可以接受的。

然后,用某个适当选择的模型对稳定性研究的整个数据组(参照所有设置水平)进行回归,为任何设置水平提供模型参数的估计。这些估计将允许 RM 生产者:

- 用检出的不稳定性估计材料的最终有效期限(在标准样品生产者场所的储存条件下);
- 估计材料稳定性的不确定度;
- 估计看似稳定的材料(证后监测)的最大允许重新检测间隔;
- 评估材料在交付给客户期间的最大允许设置负荷(温度和时间)。

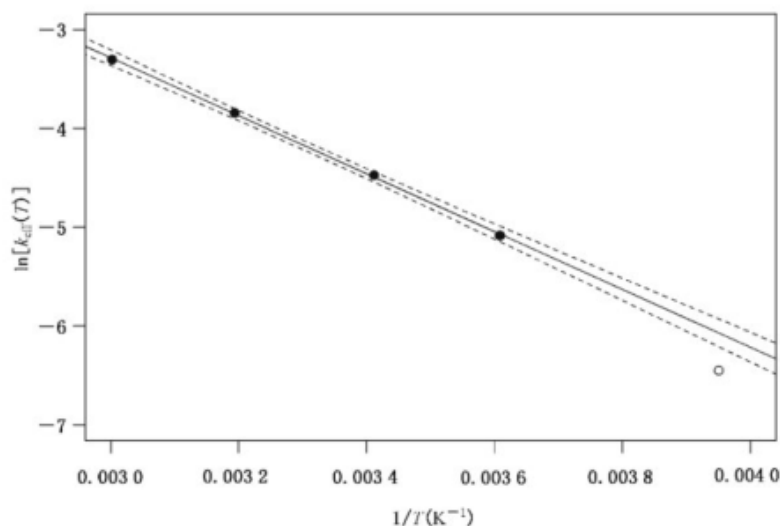
根据 8.4.3 和适当的研究设计(8.4.4),在最高温度比计划储存温度高至少 20 °C 的加速研究中,在任何设置水平下均未发现降解,计划的储存条件下的降解速率可以假设为忽略不计,但须经适当监测计划确认(见 8.10)。

超出测试储存条件范围的外推(例如,仅从温度高于 0 °C 的试验中预测 -20 °C 时的降解率)可能是不可靠的,且不推荐。

注 1: 某个材料降解/变质的过程可能很复杂,可能需要不同的模型。出于稳定性研究的目的,在许多情况下,复杂关系式 $x(t) = F(T, t)$ 被分解成以下的形式:

$$x(t) = F(T, t) = f(T)g(t)$$

示例: 图 7 显示了在环境 CRM 中试验获得的总石油烃(TPH)含量的相关关系。



注 1: 该图显示了土壤 CRM 中总石油烃(TPH)含量的阿伦尼乌斯图。该图显示有效降解速率 $k_{\text{eff}}(T)$ (实心圆圈 •) 与温度倒数的相关关系和直线的 95% 置信区间。注意: 由于技术原因,最后的数据点(空圆圈 ○)落在置信区间之外;尽管如此,回归线仍然预测了该温度下的变化,并更好地说明了置信限。拟合关系适用于前 4 个点所代表的较高温度。

图 7 加速稳定性研究的结果

注 2: 参考文献[23]和[24]给出了更多的应用实例,并进行了全面的描述。

8.5.5 稳定性研究中随机变异的其他来源

通常,稳定性研究中会有一个以上随机效应;例如,如果在每个时间点重复观测多个 RM(标准样品)单元,则原则上存在单元间项以及误差项。需要进行专门处理(理想情况下,使用统计软件),以便充分考虑这种误差结构。最常见的方法是使用极大似然法的混合效应建模,该法允许多个随机效应、计算随机效应的方差和相应固定效应(斜率和截距)所需的标准误差^[25-26]。然而,当每个 RM 单元的重复测试次数相同时,简单的处理是平均多次重复测试以给出每个 RM 单元的单个值;这可以被视为包含单个误差项的统计模型。缺点是这减少了自由度,因此增加了与回归系数相关的不确定度。

8.6 稳定性研究中发现显著趋势后的措施

如果观测到技术上的显著趋势(见 8.5.2.6),宜采取以下方法之一:

- 对观测到降解的特性值不进行定值;
- 根据模型预测,缩短标准值的有效期,确保值的变化不是技术上显著的变化;
- 估计在预期有效期内的预期降解程度,折合成为标准不确定度,并在赋值的不确定度中包含预期降解的不确定度;
- 给出标准值及其不确定度的时间函数,反映其估计趋势及其不确定度;
- 由 b)、c)和 d)中的两种或更多种的组合。

注 1: 上述 c)中的方法等同于在测量不确定度中包括一个未校正的偏差。

注 2: 上述方法 d)需要一个很好确定的降解模型。

注 3: 8.7.4 中给出了存在显著趋势时不确定度评定的进一步细节。

示例 1: 一个 CRM 对溶液中²⁴²Pu 的浓度进行了定值。由于确定的同位素放射性衰变,该浓度随着时间的推移而降低。证书陈述该标准值在一个准确的日期(进行测量的当天)是有效的,并说明了同位素的年度损失(0.000 22%/年)、半衰期和半衰期的不确定度,以便允许计算使用日期的准确值以及当时该值的不确定度。

示例 2: 发现定量微生物 RM 平均每储存一个月损失 0.05 log CFU,其损失量标准不确定度为 0.01 log CFU/月,与特性值的不确定度相比,该损失是重要的。初始特性值是在定值时获得的值。为了允许储存中的损失,证书发布了定值后随时间推移的修正值,并且在特性值声称的不确定度中包含了修正的不确定度(见 8.7.4)。

8.7 稳定性研究的不确定度评定

8.7.1 来自稳定性研究不确定度评定的一般考虑

如果有效的技术原因证明,与特性值不确定度(如小于 $u_{\text{CRM}}/3$)相比,证书有效期内的潜在变化可以忽略不计,并且得到了经验和观测的支持,则长期稳定性的不确定度就可以设置为零或者从标准值不确定度中忽略。

在其他情况下,如果稳定性数据分析产生的估计变化率通常是作为拟合模型中的系数或由该系数导出,则可以根据模型预测未来潜在的变化。这些预测可以为监测计划的选择提供信息。此外,有可能对标准值随时间推移与预测变化相关联的不确定度进行估计。如果材料在超出稳定性研究时间的信息很少,而且 RM 生产者选择采用相对较少的监测频次(如每年或更少),那么宜谨慎地估计这些不确定度,并且对于有证标准样品,要将这些包含在与标准值相关的不确定度中。

本节(8.7)提供了基本指导,包括在这些情况下要考虑的不确定度来源和评估可能随时间变化相关不确定度的一般程序。

注 1: 即使特性的(真)值保持稳定,基体降解也可能影响由某些测量程序获得的测量值。

注 2: 基于有效的技术原因、在经验和观测支持下对稳定性进行判断,包括考虑将可能变化的大小与材料预期用途所需的不确定度进行比较。

8.7.2 预测随着时间变化的不确定度来源

稳定性研究包括以下测量变异来源：

- a) 测量的重复性；
- b) 测量系统的轮间变异；
- c) 单元间不均匀性(在批定值时)。

所有测量变异的来源[以上 a) 至 c)]都对预测的不确定度有贡献,除了在同步设计中,轮间变异性仅存在于有必要进行多个测量轮的场合。如果存在这些变异来源并对研究期间测量结果的变异有贡献,而且假如拟合适当,则估计模型系数的标准误差将适当考虑这些不确定度。

任何不确定度随时间变化的估计值中均宜包含研究中随机变异性的不确定度。

如果某个预测的变化使用了几个模型系数,则估计的不确定度宜考虑所使用系数的完全协方差矩阵,因为模型系数中的误差通常是高度相关的。

此外,还存在系统效应,包括(但不限于)：

- 由时间测量或响应测量而产生的模型系数的不确定度；
- 由模型选择引起的不确定度,例如假设为线性降解、而不是指数变化或(更为严重)自动催化模型。

系统效应对时间测量和响应测量的贡献通常比随机变异产生的贡献要小得多。通常的做法是,在估计仅用于不确定度评定的预测不确定度时不包括这些贡献,尽管在证书或使用说明书中要包含某种校正时,这样做也是谨慎的。

用于从加速研究条件外推到在规定储存条件下的实时影响模型宜是有效的,并考虑该模型的不确定度。与潜在模型选择相关的不确定度通常不包括在估计的预测不确定度中。相反,可以拟合替代的候选模型,并利用它们的预测来选择监测方案。

注：8.5.5 讨论了稳定性研究中多种变差来源的处理。

8.7.3 无显著趋势时稳定性不确定度的估计

当技术上可以判断稳定性时(见 8.7.1),可以假设材料是稳定的,并且与稳定性相关的不确定度可以设置为零,同时要有及时检出意外变化的监测方案,在理想情况下,在材料使用造成不利影响之前检出。8.10 中讨论了初始监测点和后续间隔的选择。

如果没有经验支持技术上稳定的理由,且 RM 生产者选择使用更长的监测间隔时,则宜估计与可能不稳定相关的不确定度,并将其包含在与任何标准值相关的不确定度中。与可能的不稳定性相关的不确定度宜基于在某个时间预测值变化相关的不确定度,该时间等于初始监测点至该时间时颁发的任何证书的有效期限。预测宜使用一个合适的模型拟合可获得的稳定性数据,并且不确定度宜适当考虑所有相关随机变异的来源(见 8.7.2)。

对应用于几个时间点的经典稳定性研究的简单线性模型,与预测变化相关的不确定度 u_{ls} 通过公式(10)计算：

$$u_{\text{ls}} = s(b_1)(t_{\text{ml}} + t_{\text{cert}}) \dots\dots\dots (10)$$

式中, $s(b_1)$ 是估计斜率的标准误差,计算方法见附录 B.3, t_{ml} 是赋值时间与初始稳定性监测时间点的间隔, t_{cert} 是在该时间内发布的证书有效期。

有些 RM 生产者没有提供每个已颁发证书的具体到期日期,更愿意根据材料的计划寿命设定一个有效日期。在这种情况下,公式(10)中的术语 $(t_{\text{ml}} + t_{\text{cert}})$ 可以设定为计划寿命(即 t_{cert}),或者由 RM 生产者决定,设置为到第二个计划监测时间点,该时间点取决于在初始监测时间点之后能将发现的任何变化立即通知用户。

注：参考文献[18]给出了公式(10)的理由。

8.7.4 已知显著趋势时稳定性不确定度评定

如果存在已知的统计或技术上的显著趋势,如 8.6 所示,RM 生产者要提供依赖时间的标准值,或者可以提供与时间独立的标准值并增加预测不确定度。本条提供了在这两种情况下不确定度评定的进一步的细节。

当提供依赖于时间的值时,标准值的函数宜反映趋势的最佳估计,不确定度的函数宜反映趋势的不确定度(或酌情修正),同时考虑到 8.7.2 中列出的不确定度来源。

当给出的标准值与时间无关时,RM 生产者宜为允许预期的变化增加预测不确定度。选择有效期[通常如 8.7.3 中的 $(t_{ml} + t_{cert})$],估计该时间的降解程度,将其转换为标准不确定度(例如,如果观测到线性趋势,则使用矩形分布),并将该不确定度与预测变化的不确定度合并(使用不确定度传播律)。

注:对与时间无关的标准值引起的不确定度只能应用于区间的一侧,因为降解趋势仅是单向的,因此导致不对称的不确定度。

8.8 由稳定性研究估计储存寿命(“保存期”)

如果由于标准值缺乏稳定性而可能要设定可接受的变化水平,则有可能估计储存寿命,在该寿命内,预期标准值仍可以使用。相关原则在附录 B.4 中有描述。

注:例如,可以通过限值设置可接受的变化水平。

8.9 有关稳定性管理的使用说明

标准样品的使用说明宜包括使用者宜如何储存 RM 的详细信息,以确保材料在有效期内的降解不会超过规定的标准值。该信息宜明确包括打开包装之前和之后(显然,假如 RM 可以使用多于一次)的推荐储存条件,宜包括打开单元包装后任何必要的使用时间限制。使用说明可以包括在首次使用之前验证标准样品的完整性的说明,例如,这种验证可以基于包装中的包装密封检查或温度记录设备。

如果允许重复取样,使用说明宜包括防止污染的任何预防措施和 RM 单元开封后的储存方法,这对确保剩余材料仍然适合使用以及 CRM 陈述的不确定度不受影响是很有必要的。如果重复取样影响特性值,例如由于蒸发或重复冻结,这宜在证书上注明。

注:告知最终用户在远离光源、远离热源并在干燥条件下储存标准样品是良好的措施。此类说明还最大限度地减少了需要进行试验研究以验证所需存储条件的条件范围。

8.10 稳定性监测

8.10.1 监测要求

监测计划宜贯穿 RM 寿命期。如 8.1 中所述,给定 RM 在其寿命期的行为难以从典型的稳定性研究中进行可靠预测。因为行为很难预测,通常有必要监测材料的稳定性。因此,对发布后的标准样品监测是使用寿命长的标准样品的稳定性整体管理的重要组成部分。

如果使用以前 RM 生产批的稳定性经验来发布未来批的稳定性声明,包括保存期、降解率或长期储存的不确定度,RMP 宜要有支持稳定性声明的证据。在这些情况下,监测检验宜包括以前批到期日的测量。

然而,监测并不总是必要的。有些材料的使用寿命很短,并且早在有效期满之前,可以使用接近连续生产情况下的替代材料。这些材料会在任何合理监测点之前过期。因此,RM 生产者宜评估发布后监测的需求,评估宜考虑:

- 稳定性研究的持续时间和结果;
- 相关材料(或与以前批基本相同的材料)的稳定性经验;
- 材料的预期销售寿命或可用寿命;

- 有关特性的不确定度陈述是否包括对变化的补偿；
- 假设特性值保持稳定的证据的可信度，例如来自文献研究。

与相同或非常相似材料的已知降解速率相比预期寿命短的 RM，监测是不必要的。在大多数其他情况下，通常认为有些监测是必要的，并且宜在材料的寿命期内进行至少一次以确认其稳定性。

8.10.2 初始监测点和监测间隔的选择

8.10.2.1 先验信息的相关性

如果认为监测是必要的，RM 生产者宜设置一个合适的初始监测点并为进一步监测设置时间间隔。这两种选择在很大程度上均取决于长期以来在非常相似材料上获得的信息（或相反）。

8.10.2.2 具有大量先验信息的监测计划

如果从稳定性研究和/或非常相似材料的监测中获得了可靠的相关信息，涵盖的时间相当于（或长于）所讨论材料的预期寿命，则 RM 生产者可以按以下设置一个初始监测点和时间间隔：

- 类似于成功用于非常相似材料的那些初始监测点和时间间隔；或
- 基于观测到的先前研究的材料随时间的变化。

在后一种情况下，宜根据先验信息设置间隔时间，以便合理预期监测点之间的变化，并且不超过标准值不确定度的三分之一。

在评估可能的变化时，宜考虑任何试验评估的不确定度，例如，在不确定度较大时，选择较短的间隔。在评估可能随时间的变化和选择监测间隔时，也可以考虑已发表文献、经验或其他来源的相关先验信息。

8.10.2.3 无法获得大量先验信息的监测计划

8.10.2.3.1 无法获得先验信息的情况下进行监测的一般考虑

如果与所讨论材料非常相似材料的稳定性研究和/或监测的信息不足，特别是在没有类似材料的可用于监测时间超过所讨论材料稳定性研究持续时间的前提下，RM 生产者宜根据所讨论材料的稳定性研究结果设定初始监测点，并计划至少对前三个监测点进行比较频繁的监测。

在没有关于非常相似材料的扩展数据的情况下，选择初始监测点有两种基本策略：

- a) 预测可能的变化，并在对最终使用产生不利影响的任何变化之前设定初始监测点；
- b) 使用稳定性研究持续时间的简单倍数。

8.10.2.3.2 和 8.10.2.3.3 提供了这些策略的进一步细节。

在 a) 和 b) 这两种情况下，宜在审核前三个监测点的测量结果之后，设定后续监测点（即在前三个点之后）的间隔。

注：如果长期稳定性研究时间 t_{test} 比材料的预期使用寿命长，则策略 b) 不适用。

8.10.2.3.2 使用预测到的变化设置初始监测点

使用预测到的变化设置初始监测点和监测间隔[8.10.2.3.1 中策略 a)]，需要对标准值指定允差。该程序原理上很直接。首先，在赋值之后的一系列时间内，构建标准值变化的双侧置信区间。其次，以图形或数字方式确定其中一个极限与指定容差极限相交的最早点。该时间点，或者一个方便的更早时间，被作为第一个监测点。详细在附录 B.4 中给出。

注：标准样品的指定允差可以基于预期用途的考虑。对于 CRM，该间隔通常是基于扩展不确定度。例如，选择扩展不确定度三分之一的允差，标准值在第一个监测点之前移动到扩展不确定度外的风险就很低。

8.10.2.3.3 使用稳定性研究持续时间的简单倍数

该策略[8.10.2.3.1 中策略 b)]比较简单,但可能比策略 a)的间隔更短。它是基于长期稳定性研究持续时间 t_{ls} 的倍数。宜选择所用的倍数,以限制可能在第一个监测点之前影响最终用途的变化风险。

示例:下面给出了策略 b)的应用示例。

- 将第一个监测点设定在赋值日期加上 t_{ls} 和长期稳定性研究结束日期加上 t_{ls} 两个日期中的较晚者。
- 在第一个监测点后以 $2t_{ls}$ 的间隔设置两个后续的监测点。

8.10.3 稳定性监测的试验方法与评定

8.10.3.1 经典监测设计

监测经常使用经典设计。这包括在计划时间点常规储存条件下 RM 单元的测量。结果评估包括将每项监测结果(均值)与标准值进行比对,并随时间的推移,核查观测值的任何显著趋势。

经典设计的优点是简单,主要缺点是结果可能会受到测量过程中长期变异的不利影响。

8.10.3.2 评定稳定性监测结果

用于 CRM 的单次稳定性监测试验的基本评定依赖于新测量值与标准值的比较。该方法需要与 x_{mon} 和 x_{CRM} 相关联的标准不确定度 u_{mon} 和 u_{CRM} ,以及在大约 95% 置信水平下的适当的包含因子 k 。使用该方法,如果下述条件不满足,则可以作出结论有不稳定的证据。

$$|x_{CRM} - x_{mon}| \leq k \sqrt{u_{CRM}^2 + u_{mon}^2} \quad \dots\dots\dots (11)$$

如果除了标准值之外,还可获得先前对相同值的监测结果,则宜核查这些值的趋势。可以使用简单线性回归核查两个或更多个监测点的值(除了标准值之外)的趋势。如果不同点的不确定度明显不同,则可以使用权重。如果在 95% 的置信水平梯度是显著性的,则宜得出该值存在趋势的结论。

如果公式(11)中的准则没有得到满足,或者有证据表明有趋势,则表明材料的降解是显著性的,宜采取措施。可能的措施包括:

- 开展验证性研究(暂停或不暂停分发 RM);
- 停止分发并放弃该材料;
- 对该材料重新定值。

评定之后,可将新的监测数据添加到积累的材料的稳定性数据中,以便修改未来监测点或修改标准值的估计有效期。

注:符合公式(11)不能简单认为是稳定的(特别是对未来稳定性的判断),或者认为根本没有发生变化。只能得出这样的结论:这种变化(如果有),还不足以通过监测所获得的不确定度检测出来。

8.10.3.3 同步稳定性监测

经典监测的一种替代方案是使用同步试验设计的一种变体,即半连续稳定性研究。这包括将材料的某些单元储存在参考条件下,该条件通常是非常低的温度,在该条件下预期材料非常稳定。在每个监测点,在重复测量条件下(理想上),储存在常规条件下的单元和储存在参考条件下的单元一起检测。数据处理宜包括以下一个或多个步骤:

- 与储存在参考条件下的单元进行比较,进行显著性变化检验;
- 对常规储存样品与参考条件下储存样品之间的差值或比值的趋势进行检查和分析;
- 建立允许测量过程中轮间变异的模型。

注:参考文献[18]提出了一种替代方案,其中在每个监测点将新的 RM 单元放入参考温度下,这避免了未来可能在参考温度下监测大量储存 RM 单元的必要性,但代价是需要进行更复杂的数据分析。

9 材料定值

9.1 概述

本章中的指导主要用于赋予某个材料标准值(y_{char})所进行的测量。对于非标准值(或其他名称)的测定研究,也可以遵循本章中概述的原则,但通常要求不是很严格,特别是在评定测量不确定度和建立计量溯源性时(见 9.11)。

重要的是:标准值宜是真值的良好估计,而不只是总体的平均值。对于批量生产的各个单元,标准值可能是相同的,但在生产一些单个制品的情况下,则可以为每个单元分别赋值。

对于标准值,宜确定有关的定值不确定度(u_{char})。ISO 17034 要求 RM 生产者提供标准值与某个规定参考标准的计量溯源性证据。这意味着不管采用什么方法,都宜明确阐述标准值的计量溯源性。只有当所得值在声称的不确定度范围内提供了被测量值(定义)的有效估计值,且结果可溯源到同一计量参考标准时,才能实现可溯源性。国际单位制(SI)是最理想的首选计量参考标准,但其他的参考标准也可以使用。计量溯源性也可以应用于操作定义的量;通过适当的校准,确保对规定计量参考标准的可溯源性仍然至关重要。

注:9.2 提供了有关标准样品定值中建立计量溯源性的指导。

定值可以通过在一个或几个实验室采用一种或几种方法实现^[27]。ISO 17034 列出了定值的几种基本方式:

- 在单个实验室内采用单个参考测量程序(如 ISO/IEC Guide 99 中定义的);
- 在一个或多个有能力的实验室中采用两种或更多种可证明其准确度的方法对非操作定义的被测量进行定值;
- 采用一个有能力的实验室网络对操作定义的被测量进行定值;
- 由一个实验室采用单个测量程序将某个标准样品的值赋予另一个与之高度匹配的候选标准样品;
- 根据标准样品制备中所用成分的质量或体积进行定值。

本章提供了有关这些基本原则,以及协作研究的实施、纯度定值、与高度匹配 CRM 直接比对定值的指导,但目前的技术水平还不足以定性(名义)特性(如物质的身份)的定值给出详细的指南,不过本章也列出了一些通用原则。

9.2 建立计量溯源性

9.2.1 原则

计量溯源性是测量结果的一种特征。实际上,某个特性值测量结果的溯源性包含两部分:即清晰定义被测量的身份以及该被测量的特性值与规定参考标准的溯源性。因此,建立溯源性要求标准样品生产者既要提供被测特性的身份的证明,也要提供与某个适当的规定参考标准的结果比对。通过确保测量程序得到适当确认、测量仪器得到适当校准、所有测量条件(如测试材料的制备、环境条件等)得到充分控制,建立比对以提供一个可靠的结果。RM 生产者能够通过多种方式来确保比对有效,包括测量程序的确认和在受控条件下进行仪器校准,或通过使用已知特性值的材料来验证溯源性。以下各条是对这些原则的进一步指导。

注:参考文献[28]、[79]和[80]为化学测量计量溯源性提供了进一步指导。

9.2.2 计量参考标准

测量结果的溯源性通常是通过对所有相关输入量,依据合适的测量标准和/或有证标准样品进行适

当的校准来保证。量值可溯源到：

- 普遍接受的单位制[如 SI]；
- 测量标准,包括 CRM。

大多数情况下,实验室将使用可溯源到更高参考标准(如 SI)的测量标准。只要可能,就宜去尝试。如果所有其他输入因子均得到适当校准,那么通过以上测量标准校准得到的值可通过该测量标准溯源到更高的参考标准。

注 1: 典型需要校准的物品包括天平、温度计、力矩扳手、容量器具、游标卡尺、秒表等。

注 2: 在许多情况下,测量标准(包括 CRM)作为测量程序的校准标准,例如工作标准(可溯源到原级标准)或约定尺度,如 pH,其约定的基准复现是用哈内德(Harned)池,而常规校准则使用缓冲溶液。如果所有其他输入因子都已经得到适当校准,那么通过以上标准校准得到的值可通过这个标准的值溯源到更高的参考标准。

注 3: 如果用于校准的某个 CRM 标准值本身可以溯源到更高参考标准(如 SI),那么假如所有其他输入因子均得到了适当校准,新的 CRM 将通过该 CRM 的校准溯源到这个更高参考标准。

注 4: 相关输入量通常依据测量结果的合成标准不确定度进行评定。常见的经验法则是,如果一个相关输入量的不确定度大于合成标准不确定度的三分之一,则考虑该输入量的不确定度贡献。

9.2.3 被测量的类型

被测特性可能包括以下方面。

- 不涉及特定测量程序的定义。例如,基本物理特性(长度、质量)和有明确定义的、能够直接与物质的量(摩尔)建立联系的物质浓度。这种情况下,被测量与某个特定测量程序无关很重要。
- 操作定义的。这种情况下,被测量通过一个文件化且被广泛接受的测量程序来定义、且只有通过相同测量程序获得的结果才能进行比对。

无论被测量是否是操作定义的,建立溯源性均需要同样的活动;对测量结果有重大影响的每个量,均宜进行校准或保持在适当的控制之下,如通常使用校准过的仪器^[81]。

示例 1: 某个样品中 Cd 浓度进行认定。所选测量程序是基于酸消解的电感耦合等离子体质谱(ICP-MS)和中子活化分析(NAA)两种经确认的程序。测量使用有证镉溶液对测量进行校准。由两种完全不同测量原理的程序得到的结果在各自不确定度范围内取得一致。结合确认数据和校准证据,这使得确信与不确定度相比任何一种程序的偏差都很小,表明该被测量不是操作定义的、且标准值可溯源到规定的参考标准。

示例 2: 通过实验室间协作研究,其中所有参加者均采用 ISO 6865,测定 ISO 6865^[29]中定义的粗纤维质量分数。所有的测量条件(温度、体积、质量等)均得到适当校准,该被测量是操作定义的(ISO 6865),且结果可溯源至 SI。

示例 3: 使用近红外光谱(NIR)测定粗纤维的质量分数。使用依据 ISO 6865 获得的测量结果来校准仪器,该被测量是操作定义的,且量值可溯源至 ISO 6865 获得的结果。

注: 由于不确定度必须包含校准值的不确定度,以及使用 NIR 程序引入的额外不确定度(包括对特定材料可能存在的程序间偏差的补偿),因此,由 NIR 程序所得结果的不确定度通常比定义程序的更大。

9.2.4 样品制备或前处理的影响

对很多基体标准样品,情况比较复杂。尽管经校准的测量仪器测得的特性值可溯源到合适的单位,然而前处理步骤,如提取、预处理或将样品从一种物理或化学状态转移至另一种状态,却很难被校准。这些处理只能与某个参考程序(当可获得时)进行比较、或在它们之间进行比较。这使得对被测量的明确定义有些复杂。通常存在三种可能性。

- a) 对于有些处理,参考测量程序已经确定,并可用于定值研究,以提供通过参考测量程序确定的标准值。这给出了操作定义的量。
- b) 第二种可能是使用两种或更多种独立的程序评估程序的偏倚。如果由独立程序得到的结果在各自不确定度范围内一致,RM 生产者可以断定所获得的值不受各个程序的显著影响,因此该被测量不是操作定义的。
- c) 其他情况下,只能在使用相同程序的不同实验室之间进行比对。在这种情况下,不能证明不存

在方法偏倚；因此，结果是操作定义的被测量。

示例：土壤中 Cd 的质量分数，由几个实验室均采用王水提取后 ICP-MS 定量的方法进行测定。该被测量是操作定义的，即“由王水提取后，通过 ICP-MS 定量分析获得”。

在提供给用户的文件中，特性的定义宜反映所选择的定值方法。

9.2.5 溯源性的验证

很多情况下，很难证明对每台设备和设备的每个部件进行了适当校准。这种情况可能会因一些未知的影响因素而发生，然而在涉及多个实验室的定值研究中会更频繁地出现，在这种情况下获取每台使用仪器和每台仪器使用部件的校准证书是不现实的。在这些情况下，为确保设备正确校准和结果溯源性而采取措施的充分性，通过特别设计和制备的控制样品（这种样品不用于校准）与 CRM 进行验证。质量控制样品结果的一致性可以作为对所有相关输入因子充分校准的证明。

注：符合 ISO/IEC 17025 的证据，包括来自第三方评估的证据，均可以作为不同测量实验室报告结果溯源性的补充证据。

9.3 在单个实验室中采用单一参考测量程序（参见 ISO/IEC Guide 99 定义）进行定值

9.3.1 采用单一参考测量程序但无需与同类某个 CRM 直接比对进行定值

9.3.1.1 概念

在该定值方式中，由一个实验室仅使用一种测量程序赋值，而无需与某个与之高度匹配的 CRM 直接比对。由于受测量程序和实验室数量的限制，极大地制约了检出未知影响的可能性。因此，这种方式要求获得一个测量程序，该程序可以被充分理解、未知的影响可以被排除。

注：“同类 CRM”是指某个 CRM 其与待定值的 CRM 的所有特征（如基体、被测特性、被测特性的量值等）均相匹配，这些特征可能对测量结果具有某种影响。

结果及其不确定度用于确定所赋予的值。

9.3.1.2 测量程序要求

采用这种方法的所有测量程序均宜满足以下要求：

- 对测量程序完全掌握，这意味着所有的步骤都有坚实的理论基础，因此，相对于预期用途的系统误差可以忽略；
- 测量程序完全由一个包含了所有相关影响因子的测量方程来描述，这些影响因子将被测量与实际测量特性联系起来，所有这些特性都可以用 SI 单位来表示；
- 测量方程不包含对测量结果有重要影响的经验测定因子（如“回收率”）；
- 被测量对方程中所包含的所有影响因子都没有影响；
- 方程中包含的常数是已知的，而且其不确定度较低，可以用 SI 单位表示；
- 在对方程中所含影响因子分别量化的基础上，实际不确定度评定方案可以用 SI 单位来记录；
- 相对于该 RM 的预期用途，由测量程序获得结果的测量不确定度足够微小。

上述要求的建立宜通过验证，如第三方评估、适当确认研究、基于 ISO/IEC 17025 的测量不确定度评定、通过与其他实验室比对的能力确认、能力验证等予以证明。

所赋的值是由参考测量程序获得的结果，结果的标准不确定度用 u_{char} 表示。

除了使用参考测量程序测量之外，还强烈推荐使用另一个独立测量程序进行测量确认，以避免粗大误差。严格来说，因为确认测量结果的不确定度高于参考测量程序的不确定度，即使通过另外的独立测量程序进行结果确认并非必要，但为了进一步增强对结果的信心，仍强烈建议这样操作。

确认测量的结果不必与参考测量程序的结果进行合成，因为它们的不确定度通常要高得多。相

反,需要对不同程序的结果进行检验,以确定独立程序的结果与参考测量程序的结果是否一致。如果一致,就不存在方法偏差的证据。如果不是这种情况,则需识别原因(在确认程序中出现的偏差或在参考测量程序中出现的意外影响),如有必要,还需对结果进行校正。

9.3.2 由一个实验室采用单一测量程序,将标准样品的值传递给与之高度匹配的候选标准样品

9.3.2.1 原则

在该定值方式中,通过将候选 CRM 的结果与已定值且与之相似的 CRM(“原级 CRM”)的结果直接比对赋值于“次级 CRM”。这类材料的例子包括:用有证溶液测量比对的痕量元素溶液、用药典标准测量比对的材料、或用有证吸光度标准测量比对的吸光度标准。

注 1: 对候选 CRM 的每项测量都要求校准,实际上就是与另一个 CRM(校准物)进行比对。本条款完全针对这种情况,即两种 CRM 如此相似,使得在一个实验室用一种方法通过直接比对就可以赋予标准值。若使用不同类型的 CRM(如使用纯溶液校准基体材料的测量),则属于另一种定值方法。

对于该方式,次级 CRM 宜与原级 CRM 在所有对测量结果有重要影响的特征方面都充分相似,从而使得在陈述的不确定度范围内,次级 CRM 因材料引起的偏差可忽略不计。对此,宜考虑以下方面:

- a) 原级和次级 CRM 由相同的基体组成。只要材料的主要特征保持不变,可以接受在建立校准曲线时存在的微小的差异。原级和次级 CRM 对于所使用的方法面临相同的分析挑战。

示例 1: 能够使用有证 HNO_3 中 Cd 溶液对 HNO_3 中 Cd 溶液进行定值,因为两个 CRM 的基体(稀释的 HNO_3)是相同的。但不能使用 HNO_3 中 Cd 溶液对作为次级 CRM 的花岗岩中 Cd 进行定值,因为基体不同。对很多方法,消解步骤同样也增加了额外分析的挑战。

示例 2: 如果在次级 CRM 色谱图中出现了与目标分析物一起的共流出物,就不可能将其当作次级 CRM 进行定值。如果原级和次级 CRM 的差异比较复杂(如使用一系列未混合的单组分原级 CRM 对多组分次级 CRM 进行定值),那么 RM 生产者宜证明这种增加的复杂性不影响该结果。

- b) 如果被测量不是操作定义的,对所讨论的测量,基体是某种被视为完全了解的测量程序。

示例 3: 充分了解色谱测定苯并[a]芘溶液,因为没有共流出物或基体效应发生。但不完全了解测定土壤(或在土壤提取液)中苯并[a]芘,因为很多因子会影响提取效率,且可能会产生很多共流出物。

- c) 被测特性含量水平的差异不会导致原级和次级 CRM 测量结果之间的显著偏倚。

注 2: 对于化学测量,这些条件实际上将次级 CRM 的生产限制在纯物质、纯物质的溶液/稀释液或操作定义的特性。

定值所用测量程序宜满足 ISO 17034 中列出的所有溯源性准则,并说明原级 CRM 定值的被测量。

RMP 宜证明将原级转移到次级 CRM 的值和不确定度的有效性。

9.3.2.2 赋值和 u_{char}

赋值通过直接比对原级和次级 CRM 之间的结果由计算得到。有效的方法包括插入法、使用原级 CRM 作多点校准曲线、使用高度匹配标准值的原级 CRM 作单点校准并将测量差异加入标准值中。

u_{char} 由原级 CRM 标准值的不确定度、根据选择的校准模型(包括技术选择性的贡献)校准的不确定度,以及重复性对次级 CRM 结果的影响的合成构成。计算的不确定度宜考虑用于赋值的特定统计处理。

9.3.2.3 溯源性

次级 CRM 的标准值可通过原级 CRM 溯源到与原级 CRM 的值相同的参考标准。

示例: 使用有证 HNO_3 中 Cd 溶液(原级 CRM)对 HNO_3 中 Cd 溶液(次级 CRM)进行定值。原级 CRM 的标准值能够溯源到国际单位制,因此,次级 CRM 的标准值也能溯源到国际单位制。

9.3.3 单个实验室定值时 RM 单元的选择

RM 生产者宜制定测量方案(RM 单元的数量、重复结果的数量,等),该方案能够实现每个标准值

的预期不确定度。

9.3.4 配方方法

这种方法通常适用于使用纯物质生产校准溶液,也用于混合气体,关于混合气体的生产在另一个标准中有介绍^[30]。这种方法有时也用于基体材料的生产。

为了计算标准值和不确定度,必须了解用于混合的所有材料的被测量的值及其不确定度。在很多情况下,这相当于测定目标材料的纯度(见 9.6),以及确认其所添加的材料中不存在目标材料(如溶剂或“空白”基体材料)。

重要的是要防止材料在获取和混合时含量发生变化;例如,只要适用,宜排除水分的损失或吸收。

如果采用重量法制备几种材料的混合物,且所有这些材料均包含所讨论的被测量,则每种材料均宜使用本条款中描述的方法之一进行定值。

容量法生产遵循相似的原则来计算所赋予的值及其不确定度,但需特别关注混合液的体积不可简单加和的情况(如乙醇/水的混合物体积并不是将要混合的水和乙醇的体积进行简单加和)和其他影响体积测量的因素(特别是温度)。

对于采用纯重量法生产定值质量分数的情形,所赋予的值 y_{char} 由各个组分的质量 m_i 和每种材料的质量分数 w_i 计算得到,如公式(12):

$$y_{\text{char}} = \frac{\sum w_i \cdot m_i}{\sum m_i} \quad \dots\dots\dots (12)$$

对于相同的程序,不确定度 u_{char} 包含所有称量步骤的不确定度以及各个质量分数的不确定度。对于最常规的两种物质的混合, u_{char} 可以采用公式(13)计算:

$$u_{\text{char}} = \sqrt{\left[\frac{m_1}{m_1 + m_2} \right]^2 u_{w_1}^2 + \left[\frac{m_2(w_1 - w_2)}{(m_1 + m_2)^2} \right]^2 u_{m_1}^2 + \left[\frac{m_2}{m_1 + m_2} \right]^2 u_{w_2}^2 + \left[\frac{m_1(w_2 - w_1)}{(m_1 + m_2)^2} \right]^2 u_{m_2}^2} \quad \dots\dots\dots (13)$$

注 1: 公式(13)不包含称量或 w_i 的测定过程中相关误差的协方差项。正相关影响会导致合成不确定度变大。

原则上采用重量法配制足以进行赋值,但仍强烈建议通过测量来核查配制过程中的任何错误。

注 2: 关于气体分析中重量法的使用在文献中有详细讨论(如参考文献[31]和[32])。

9.4 在一个或多个有能力的实验室采用两种或两种以上可证明准确度的方法对非操作定义的被测量进行定值

9.4.1 概念

9.4.1.1 对于很多被测量,还没有能在适当不确定度水平下提供准确结果的参考测量程序。在这些情况下,有必要寻找其他方法来提高赋值的可靠性。本条中所述的方法使用许多由不同测量程序和/或从不同实验室获得的数据组,以便:

- 通过表明各独立程序产生相同的结果,证明测量程序中没有显著的偏倚;
- 通过结果之间的一致性,证明每个实验室没有显著的实验室偏倚;
- 通过平均结果提高赋值的可靠性,从而降低重复性和随机性的影响,并降低实验室间或方法间变异的影响。

9.4.1.2 根据不同测量程序(可能在不同实验室执行)结果的一致性,测定独立于方法的 RM 特性值的概念基于至少两个假设:

- 存在多个有能力的程序和/或实验室,可测定 RM 特性并提供准确度满足要求的结果;
- 对大多数数据评估方法,假设不管任何原因(如测量程序、人员或仪器的变异),在测量程序/实验室内和间各个结果之间的差异本质上是随机的。

9.4.1.3 为了使这种方法有效,所有涉及的测量程序和/或实验室都宜测量相同的被测量,且测量结果

宜溯源(见 9.2.2)到相同的单位制。这需要仔细选择校准标准、并仔细调查所使用的测量程序。9.4.2 描述了校准标准的选择。

注 1: 即使处在最高水平,实验室之间也可能存在测量程序性能特征的差异和不确定度大小的差异。

注 2: 实验室间比对的目的可能是不一样的,其中包括方法确认、能力验证和标准样品定值。本研究的目的对于各种研究的建立和评估具有重要的意义。因此,重要的是要牢记定值的目的,即使它们在逻辑上相关联,也不要将其与其他目的混淆,如能力验证演练(见附录 A.3)。

9.4.1.4 实验室间和多种方法定值在一定程度上取决于不同偏倚来源的平均化,以降低不确定度。有效的平均取决于对不同影响的代表性抽样。这对于选择参加者和测量程序具有重要的意义:

- 只要可能,宜选择能很好代表不同测量原理的测量程序;
- 参加者的选择宜代表有能力的实验室。

测量程序的选择在附录 A.1.1 中讨论。具有代表性参加者的选择在附录 A.1.3 中讨论。

9.4.2 研究设计

在一个涉及多种方法的研究中,宜包含至少两种完全不同测量原理的方法。对于由许多参加者自行选择测量程序的实验室间研究,宜寻求适合测定特定特性的具有良好代表性的测量程序。

宜考虑校准标准的选择,即:每个参加者是否宜使用各自所选的标准,或是否为所有参加者提供相同的校准样品。还宜适当考虑所使用校准样品的纯度。

实验室的选择宜基于已证明的能力。参加实验室最好在研究开始前,提供与候选 CRM 测量无关的被测量的能力证明。因此,不能将来自同一研究中候选 CRM 的数据用作能力证明并给 CRM 赋值(如使用能力验证结果的公议值给 CRM 赋值)(见附录 A.3)。

RM 生产者宜为考虑赋值设定一个记录在案、技术有效的最小结果数量。数据组的数量宜足够多,以便考虑到报告失败的可能性、使得因技术原因和统计评价剔除若干结果之后,仍可对估计值提供适合用途的不确定度。

注 1: 独立数据组的数量比参加实验室的数量重要。单个实验室有可能通过独立的测量程序、校准样品和/或仪器得到几个数据组。

组织者宜采取适当的质量控制措施,以确保所提交结果的质量。

生产者宜规定报告的形式,规范宜包括:

- 报告独立观测结果、平均值或两者的说明;
- 定量结果的测量单位;
- 定量结果的有效数字的位数;
- 只要适用,测量不确定度所要求的形式;
- RM 生产者要求的其他信息的性质和形式(如使用的测量程序和测量标准、测量日期和时间、或测量轮顺序)。

报告可以由每个重复测量的各个结果(带有或没有不确定度)组成、或由具有陈述不确定度的单一结果组成,这导致了不同的审查方法和评价(见附录 A)。

报告形式还包括为参加者预先格式化的报告表格。报告宜足够详细,以便于核查结果的技术有效性,包括溯源性信息。

注 2: 当报告以电子表格形式提交时,通常可以通过使用电子表格软件中的“锁定”或“保护”功能来避免无意地更改。

组织者宜为参加实验室和/或操作者提供足够的指导,以确保工作顺利进行。为了获得成功,实验室间研究宜有明确的目标,有效的设计和组织的,所有参加者易于执行的简明、清晰的指南。无论是操作者还是实验室,参加此类项目就意味着同意遵守这些指南。

注 3: 关于组织一个或多个实验室进行多种方法研究的其他指导在附录 A 中给出,附录 A 是本条的扩展。

9.4.3 评价

9.4.3.1 技术和统计评价

数据组宜进行目视和图示检查。宜核查每个实验室提交数据的完整性,发现任何异常均宜仔细检查可能存在的小问题(如传输错误、印刷错误等)和大问题(如中途退出、设备故障等)的原因。如果怀疑有转录错误,宜联系可疑实验室查询所报告的数值,但此时不宜给出预期值。如果确认存在错误或故障,则相应的结果宜予以校正或删除。

根据研究参加者提供的测量程序信息,核查所有结果是否有技术错误的证据。技术评价宜得出一组技术上有效的数据,即每个单独获取的数据都将被视为真值的无偏估计。

注1:术语技术错误指的是根据科学证据从数据组中排除的测量结果,该术语不是仅指基于统计考虑而表明与数据组离群的测量数据。

注2:在这类研究中,已发现插入已知特性值的质量控制样品有助于识别技术问题。

宜对技术上接受的所有数据组进行统计评价,并要充分考虑到组间差异(特别是方法间和实验室间的差异)、值的可能分布、出现结果聚类 and 潜在离群值的证据。针对数据组和定值特性宜选择适当的统计方法。

当生产者要求报告测量不确定度时,技术和统计审查时也宜考虑所有报告的不确定度信息的有效性,结论宜考虑所报告的测量不确定度。

9.4.3.2 赋值和不确定度

赋值宜使用合适的统计程序,所使用的程序对特定数据组宜是有效的。

注:统计程序的确认可以包括有可靠理论基础的证据(通常通过参考合适的文献),在预期使用条件下的已知性能,以及可以证明足以应用于现有数据的假设或条件。

平均和加权平均两种常用程序的使用说明见附录 A.2.4。

定值不确定度 u_{char} 可以通过使用实验室提交的不确定度陈述或仅从提交的数据(不考虑实验室不确定度陈述)来估算,或根据两者的合成来估算。更多的信息见附录 A.2.5。

9.4.4 单一实验室多种方法研究

在某些情况下,组织方在方法开发方面投入了大量的精力,使得测量程序的计量控制接近参考测量程序。在这种情况下,只要测量程序的测量原理足够不同,由这些测量程序中少数几个程序获得的数据组就足以进行定值。在下列情况下:

- RMP 有可能获得完整的质量保证和确认数据,在进行技术评价时宜考虑这些数据;
- 数据组的数量很少。因此,宜更加重视测量不确定度的评估和正确处理。评价宜依赖于与每个测量程序相关的测量不确定度的评估和利用。

如果结果在所陈述的不确定度范围内一致,则可使用加权平均(附录 A.2.4)及其相应的不确定度。如果有效结果在所陈述的不确定度范围内明显不一致时,则宜重新慎重考虑测量程序的计量控制是否确实满足这种方法。如果确定满足,则宜在标准值不确定度中允许结果过度离散的影响。

注:补偿结果过度离散的方法包括曼德尔(Mandel)和葆拉(Paule)^[38]、范格尔(Vangel)和鲁金(Ruhkin)、伯奇(Birge)等。详情可参阅参考文献[34]。

9.5 采用有能力的实验室网络对操作定义的被测量进行定值

9.5.1 概念

该方法适用于被测量由操作定义的 RM 的生产。在这种情况下,测量程序定义了被测量,通常只

能通过合成几个实验室的数据,才可能证明没有实验室偏倚。另外,定义程序通常是相对不精密的,获得小不确定度的唯一实用方法是对来自不同实验室的许多结果进行平均。

除了所有实验室均采用相同的程序之外,该方法与 9.4 中描述的方法基本类似。

假设仍然存在一些实验室,它们能同样出色地进行所讨论的测量。该方法旨在将测量程序设定范围内的所有影响因子随机化。

9.5.2 研究要求

宜选择很好描述的测量程序,这宜是已出版的标准方法,最好是国际上商定的程序(如 ISO、ASTM、AOAC、IFCC 等)。宜要求参加者严格遵守程序,且只准许程序中允许出现的变化。

注 1: 所有参加者同意的对此类程序的任何修改(如对某些参数的更严格规范)原则上导致修改程序,然后参考修改后的程序来定义被测量的身份。

注 2: 初步研究表明倘若意外偏离了标准程序,在进行定值之前可对其进行纠正,以确保遵循标准程序。

注 3: 在这种情况下,质量控制样品也可用来证明某种特定仪器是否满足所有规范。

9.5.3 评定

对于操作定义的被测量,定义程序是(根据定义)无偏的,因此,只要在不确定度评定中考虑可能的实验室偏倚和实验室内影响。

附录 A.2 中描述的方法适用于对结果评价。

9.6 纯度

9.6.1 概述

一般而言,纯物质是化学、测温学、量热学中大多数溯源链的原级测量标准和更高级计量溯源性的最终来源,尤其是溶液和基体标准样品的定值。形容词“纯”是指一种理想化的情况:没有任何纯物质是 100% 纯的,在某种程度上总会有杂质存在。因此,物质纯度的适当认定是化学测量溯源性的重要基石。

纯物质本身就是一种重要的 CRM。它们被实验室用来将更高级溯源性传播到测量程序中使用的校准标准,或用于其他 CRM(如溶液或气体混合物)的认定和生产。

物质的纯度可以通过测量有关物质的量直接测定,也可以从 100% 中减去所有杂质的质量或摩尔分数间接测定。

在对某种材料的纯度定值时,还宜确认该材料的身份。

9.6.2 纯度的直接测定

在某些情况下,可以直接测定有关物质的质量或摩尔分数,合适的方法包括库仑法、滴定法和量热法(凝固点降低法)。对于有机分析物,越来越多地采用定量核磁共振技术直接定值标准样品的纯度^[39-40]。

原则上,需要用有关物质进行校准的方法(如 HPLC、GC、ICP-MS 或 AAS)可用于纯度赋值,但它们是次级测量程序。由于它们需要使用已知纯度的标准,因此,应用这些方法直接测定纯度通常限于给工作标准赋值。

以所用程序测定的结果为纯度赋值,并且以纯度测定的不确定度作为 u_{char} 。

强烈建议通过独立测量确认这些值。

9.6.3 纯度的间接测定

纯度可以通过使用一组能够检出和量化材料中所有主要杂质的正交分析技术进行差减测定,如下所示:

- a) 研究一系列可能存在的杂质,通常包括残余有机溶剂、水分、无机和有机杂质,被研究杂质的种类通常可以从物质的生产过程中了解到;
- b) 对待定值物质中所有可能存在的杂质进行测定;
- c) 主成分的纯度通过差减法计算。

测定杂质所必需的测量可能具有挑战性,因为大多数杂质接近检出和/或测定限,或者杂质与主成分在结构上很相似以至难以分解。而且,不同的测量程序得到的结果的单位不同(如挥发物为质量分数、用量热法测得的总杂质为摩尔分数),如果不知道杂质的结构,那么从严格的计量意义上说,就不能将这些结果合并。根据特定的校准样品对每种杂质进行量化是一种理想情形,但是如果没有足够的资源或合适的标准样品,则可能是不切实际或不可能的。在这种情况下,必须对作为对各个杂质的身份和响应因子假设的结果引入的不确定度给予适当补偿^[41]。

注: ICH 协调的三方指南 Q3A“新药中的杂质”^[42]要求对超过 0.05% 的杂质(取决于每日摄入量)进行鉴定。

尽管对各个杂质量化会获得较高的相对不确定度,但只要其绝对水平很低,对主成分最终值的不确定度的贡献通常较小。

主成分 y 的物质的量或质量分数的标准值 y_{char} 与具有物质的量或质量分数 w_i 的杂质 k_i 的函数模型由公式(14)给出:

$$y_{\text{char}} = 1 - \sum w_i \quad \dots\dots\dots (14)$$

假设各杂质质量分数的测量是相互独立的,那么主成分的物质的量或质量分数相关的合成标准不确定度为:

$$u_{\text{char}}^2 = \sum u^2(w_i) \quad \dots\dots\dots (15)$$

式中, $u(w_i)$ 是 w_i 的标准不确定度。由于有些杂质确实不存在,或者它们的含量低于测量程序的检出限,因此经常会出现一些杂质的质量或质量分数 w_i 为零的情况。当一种杂质含量低于检出限时,那么其 w_i 就可设为零、或赋予另一个值(通常与检出限相关)并具有相关的不确定度。

由于接近物理极限(物质的量和质量分数仅定义在 0 和 1 之间),不确定度的评定也会变得复杂,这可能会产生其他问题,包括对一些有贡献杂质种类(其中包含名义上的负值)的估计量(见参考文献[51])。

9.7 身份

9.7.1 依赖来源认定材料

标准样品可以根据材料来源的知识(即材料的来源)进行定值。

为了支持基于来源的定值, RM 生产者宜获得材料来源的文件或其他证据,提供从来源到最终包装的完整证据链,并在寿命期内保持有效。

RM 生产者宜有适当的程序来确保材料在处理过程中(包括取样、均匀化、包装、储存等)免受其他材料的污染,且不改变该材料对预期使用典型检测方法的响应。

只要可能, RM 生产者宜进行试验验证(包括测量、专家检查或定性检测,可酌情而定),以利用来源确认所赋予的身份。

示例: 从单一细菌培养得到的培养物中提取的 DNA, 该单一细菌的菌株又是从参考菌株库中分离得到的, 可以根据来源对 DNA 进行定值, 并进行污染确认核查^[43]。

9.7.2 依赖身份测量认定材料

9.7.2.1 通则

当根据测量对物质的身份进行定值时, 宜考虑以下几个方面。

- a) 身份通常不是某一个测量结果, 而是由一种或几种方法的测量结果得出的结论。例如, 核磁共振谱中化学位移和峰高, 或者与颜色、熔点、摩尔质量等的组合, 可以得到赋予该物质的身份。

而当测量不确定度可以赋予各个测量结果时,并不能直接将它们合成起来以给出身份中不确定度的任何数值指示(例如,赋予的身份正确的概率)。

示例 1:通过 DNA 测序鉴定来说明身份的不确定度和与测量结果相关联的不确定度之间的差异。DNA 序列是测序试验的结果,测序中碱基对和其他错误发生的概率,以及不同 DNA 分子所存在的突变差异原则上都可以被估计。然而,在与参考序列的同源性百分比相对较低的情况下,通常可以相当有把握地确定生物物种的身份。因此,个别测序错误可能不会明显影响到身份的赋予。

- b) 只有当结论的错误概率可忽略不计时,认定身份的 CRM 在实践中才有用。
- c) 材料的细微不均匀性和不稳定性不一定改变身份的结论。评估均匀性和稳定性的指导原则是材料的适用性,即它是否仍然支持确切的身份鉴别。
- d) 对于所选的鉴别方法,不同的物质可以具有相同的特性。因此,有关原料来源和待定值材料的加工过程的信息对于身份认定至关重要。
- e) 与任何材料一样,项目策划宜根据材料的预期用途对身份信息的需求进行明确定义。

示例 2:对于 DNA,预期用途只需要有关物种身份的一个陈述、一个完整的序列、或关于甲基化程度的附加信息。

注:身份有时由专家判断决定(如石棉纤维或微生物物种)。然而,这种判断通常是基于观测和与规范的比对。基于观测的专家判断属于本条款的范围。

9.7.2.2 规范

某种材料的身份测试涉及将该材料的一组测量结果与这些测量结果对应的规格进行比对(如熔点范围、与参考 DNA 序列同源性的百分比)。

示例:有机高分子材料可根据与参考红外光谱的比对,使用下列准则进行鉴定:

- 与参考谱图中的所有峰值频率偏差在 3 cm^{-1} 以内;
- 与参考谱图的相对峰强度偏差在 5% 吸光度范围内;
- 参考谱图中有的峰均无缺失;
- 候选 RM 谱图中的所有峰在参考谱图中均有。

规格的来源可包括国际上认可的纲要(如药典来源[44])和其他参考资料集(如[45])。这些信息可以在 RM 生产者控制之外发生变化。因此,RM 生产者宜明确说明用于物类鉴定的规格,既可以是一组数值、也可以是一个在证书上注有日期对外部规范的引用。

在编制规范时,RM 生产者宜比较各种文献资料,建立报告值的范围,并建立和记录每个被测量的规格,以反映所使用信息的范围和可靠性。宜优先考虑已经过同行评审的参考资料。

9.7.2.3 通过方法组合进行身份定值

该方法特别适用于确定的中、小分子质量的化学物质。

宜选择多种方法探测候选标准样品的不同特性。通常使用的方法包括,如熔点、摩尔质量、UV、IR、NMR 和质谱的测定。结合有关原料及其加工步骤的信息,以及取样和运输到 RM 生产者的信息,收集的方法宜毫无疑问地足以确定材料的身份。如果存在详细的、已发布的规范(如用于鉴定的药典标准),则方法的选择可能仅限于这些规范中所列的方法。

注 1:确定身份所需方法的性质和数量会随着潜在的类似产品的数量(如有机物质比无机物质更多)和有关来源和加工步骤的信息而变化。

所有使用的测试和测量程序均宜经过适当确认,且结果宜符合 9.2 中有关溯源性的要求。如果可能,在定值过程中宜与 RM 一起检测适当的控制材料。

每项测试和测量所得的结果均宜遵循已发布的此类比对程序,与拟用的规格要求进行比较。若不存在此类程序,在考虑测量不确定度的情况下,测量结果与规定值不宜有差异。

若结果与规格要求一致,则可以确定身份,不确定度忽略不计。

注 2:对累积的测量和来源信息是否足以毫无疑问地确定身份的判断多少有些主观。因此,要大力提倡 RM 生产者建立同行审查制度。

9.8 存在/不存在

存在/不存在是以定性的方式评估定量测量的一个例子。高于预定阈值的结果归类为“存在”，低于预定阈值的结果归类为“不存在”。

注 1：很多测量以存在/不存在来评估，却从未被量化。然而，即使采用这些方法，通常有一个响应限值表示“存在”。这个限制的存在表明，尽管用于定性评价，该测量本质上是定量的。

对测量进行定量评价是解决这类问题的一种方法。尽管测量不确定度通常很高（这是定性评价的原因），但这种方法具有概念上简明的优点。最简单的情况是确定材料中不存在某物质，且存在用于该物质检测的定量方法（如食品中的某种污染物）。在这种情况下，对该材料的所有测量结果均宜低于表明物质存在的临界值，标准值表示为“ $<L_d$ ”，其中 L_d 为检出限。如果检测结果不是定量的，所有测量均宜得出“不存在”的结果，以认定该材料中没有某种物质，在这种情况下，参考值表示为“不存在”。此外，还宜给出测量程序的检出限。

注 2：术语“检出限”的 IUPAC 定义为“对于给定的分析程序，可以合理确定地检测到的最小量^[46]”。这是指当某种物质存在时可以合理确信检出该物质的（真实）浓度，并对应于 CC β （欧盟委员会决议 2002/657EC^[47]）以及“净状态变量的最小检出值”（ISO 11843-1^[48]中定义）。对于某个给定的程序，该检出限取决于重复测量的次数。

对于 CRM，不确定度陈述宜说明浓度上限的置信水平。

如果使用了几种测量程序，且所有结果均一致，那么最灵敏程序的检出限就可以作为标准值。

示例：三种不同测量程序得出的结果分别为 $<2\text{ mg/kg}$ 、 $<5\text{ mg/kg}$ 和 $<4\text{ mg/kg}$ 。由于所有结果均一致，因此，标准值被设为 $<2\text{ mg/kg}$ 。有关示例见参考文献[49]。

注 3：建议使用不同的程序和/或实验室，以避免将分析过程中的丢失误认为不存在。此外，包含材料加工过程的信息也可以支持某种物质不存在的表述。

注 4：如果某种物质的测量结果高于临界值，即便其低于定量限，仍有必要声明这种物质是存在的。更多指南见 ISO 11843-1^[48]。

9.9 定序尺度

一些特性以定序尺度表示，通常将物品放置在有序类别中。例如莫氏硬度标度，或将皮肤刺激分为无反应/中度发红/显著刺激/强烈反应。这些尺度通常是通过参照特定的分类方法来定义。因此，在这种情况下，唯一可能是几个实验室使用相同的方法进行定值。

在很多情况下，某种 RM 只有在没有任何争议的情况下判定为一个类别才有用。为了实现这个目的，由所有参加实验室获得的所有技术上接受的测量都宜将该材料判定为相同的类别。

如果某些结果有分歧，且分歧不能由技术错误来解释时，就不能可靠地赋予类别。然而，可以将技术上有效结果的中位数和/或众数作为信息值给出。

9.10 定性特性

可以对材料的定性特性，如颜色、气味、形状进行表征^[82]。在某些情况下，这些特性可被量化，且在实践中经常使用这种定量形式。例如，粒子的形状参数或亨特（Hunter）体系中的颜色^[50]。这就将问题转化为如上所述的方法定义被测量的定值问题。特别是颜色，也可以考虑采用吸收/反射光谱定值。

9.11 非标准值的定值

根据“标准样品”和“有证标准样品”的定义，只有标准值才需要有测量不确定度和计量溯源性陈述。除标准值之外，也可能会赋予非标准值（如“指示值”“信息值”或“参考值”），不过这些非标准值不能作为计量溯源链中的参照对象。

由于对这些值的不确定度和溯源性没有要求（即不要求可比性），因此可以使用更广泛的方法进行赋值，包括使用关于典型特性的文献数据或来自单个实验室特定条件下的数据或来自几个实验室的合

并数据。选择的表征方法越接近标准值的表征方法,所赋予的非标准值就会越可靠。

建议提供关于非标准值来源的信息,以及不认定为标准值的原因,以允许用户评估其适用性。虽然不作要求,但关于溯源性和不确定度的描述可增加非标准值的可用性。

10 测量不确定度评定

10.1 评定 CRM 特性值不确定度的基础

本文件中标准值的不确定度评定的基本原则是 GUM^[4]中规定的通用原则。GUM 提供了合成不同不确定度的一种程序,合成时每个不确定度分量均以标准不确定度表示。附录 D 给出了该程序的概要。

如果能建立完整的测量模型,并且所有不确定度分量来源均与模型中的影响量相关,则可以使用该详细程序。然而在 RM 生产中,在很多情况下,某些或所有主要不确定度均与测量输入量不相关,适合采用 10.2 给出的简化模型。

10.2 批定值的基本模型

从原理上讲,当 RM 传递给用户时,RM 单个单元中认定特性的值可能要受到来自定值过程、独立单元间实际存在的变异(非均匀性)、随时间的变化、运输和后续储存期间的变化等影响。用于评定与标准值相关的不确定度模型,当这些影响显著时,宜要考虑所有这些影响。为此,一个方便的简单模型如下:

$$x_{\text{CRM}} = y_{\text{char}} + \delta_{\text{hom}} + \delta_{\text{its}} \quad \dots\dots\dots (16)$$

式中:

x_{CRM} ——特性值;

y_{char} ——由该批定值得到的特性值或单个制品定值得到的特性值;

δ_{hom} ——不均匀性误差项,包含单元间变异以及单元内不均匀性的必要补偿;

δ_{its} ——储存条件下稳定性影响的误差项。

通常,均匀性和稳定性的研究都是假设这些误差项为零,但它们的不确定度可能不是零。

公式(16)是根据 GUM 原则建立的一个简单加和的“测量模型”。假设各变量独立,根据 GUM 中不确定度的传递规律^[4],CRM 的特性值不确定度可以表示为:

$$u_{\text{CRM}} = \sqrt{u_{\text{char}}^2 + u_{\text{hom}}^2 + u_{\text{its}}^2} \quad \dots\dots\dots (17)$$

本文件中 7.11(评定 u_{hom})和 8.7(评定 u_{its})包含了这些不确定度分量的评定,第 9 章(定值)是对特性值定值中产生的不确定度分量 u_{char} 评定的附加指南。

注 1:在某些情况下,可能要在公式(16)和公式(17)中引入附加术语。例如,运输过程不能很好控制,不能保证发生的变化可忽略,因此,需要将这种影响 δ_{trn} 和与之相关的不确定度 u_{trn} 也加到给定的模型中。

注 2:CRM 特性值的合成标准不确定度与该材料的有效期有关(见 8.6)。

注 3:有时,特性值是时间的函数,例如放射性同位素标准样品的标准值。在这些情况下,ISO 17034 要求在陈述的不确定度中适当考虑使用之前值可能发生的变化,或者如果可以预测随时间的变化,对预期随时间的变化提供一种校正标准值及其不确定度的方法。

10.3 不确定度来源

定值不确定度 u_{char} 宜考虑所有在定值测量程序中遇到的相关的不确定度来源。GUM 和参考文献[51]均列出了常见的不确定度来源。通常,测量程序已经根据测量不确定度进行评定,由此得到的不确定度评定可以应用于评定 CRM 特性值的不确定度。

通常,测量程序的不确定度模型包含组合分量,即依赖于其他几个分量的不确定度分量。在该测量程序用于常规测量时,即使这些组合分量不出现,但它们会产生协方差(见 10.1)。要获得与 CRM 特性

值相关合成标准不确定度的正确估计,对协方差和相关性评定至关重要。为了便于检出协方差,推荐用文件说明组合不确定度分量中包含了哪些不确定度分量。该文件有利于快速识别协方差和相关性的可能来源。在 GUM(ISO/IEC Guide 98-3:2008,附录 F[4])中,给出了如何评定所得到的协方差的进一步指导。

对某个具体测量程序的任何变更均宜伴随不确定度模型的评审。

10.4 包含区间和包含因子

在计算公式(17)中标准不确定度 u_{CRM} 之后,通常有必要计算扩展不确定度 u_{CRM} ,以便提供一个区间 $x_{\text{CRM}} \pm U_{\text{CRM}}$,该区间包括了大部分可合理归属于被认定特性的值。扩展不确定度使用公式(18)计算:

$$U_{\text{CRM}} = k u_{\text{CRM}} \quad \dots\dots\dots (18)$$

式中, k 是包含因子。

所使用的包含因子是根据假设的分布函数和所要求的包含概率来确定的。在许多情况下,可以假设为近似正态分布并要求的包含概率为 95%,这样得到包含因子 $k=2$ 。

当与 u_{CRM} 相关的有效自由度较小(小于 10)时,宜使用 t 分布,替代指定的 2 包含因子(2)。

如果特性值的分布是不对称的,例如服从泊松分布的计数结果,则宜说明置信区间(特定概率,即 95%),而不是扩展不确定度和包含因子。

注: 非对称分布的包含区间通常是不对称的。

附录 A

(资料性)

定值研究的设计与评价

——在一个或多个有能力的实验室中使用两种或多种可证明准确度的方法
对不依赖于方法的被测量进行定值

A.1 研究设计

A.1.1 测量程序的选择

在选择测量程序时,宜考虑以下各方面的变异:

- 样品制备,如研磨/铣削、提取或净化步骤;
- 样品导入和/或分离,如使用液相色谱或气相色谱;
- 定量方法,如分子或原子吸收、质谱、火焰离子化或荧光等;
- 校准程序,除非一种方法有明显的优点,由于其计量严谨或者由于具有较低的测量不确定度。

在许多情况下,考虑所有方面的变异是不可能的。在这种情况下,宜寻求可能的最大变异。例如,如果气相色谱是唯一可用的分离技术,那么研究至少宜包括使用不同的注射技术、不同的色谱柱和温度程序以及使用不同的检测器进行定量测试。

RM 生产者宜要求在研究中使用的所有测量程序得到适当确认,并能提供对测量不确定度的合理估计。

为了使实验室在确保必要变异的同时能够自由选择测量程序, RM 生产者宜在研究开始前获取参加者所采用测量程序的信息,以获得必要的程序范围。如果需要,宜针对提供特定方法的实验室进行有针对性的搜索,以避免仅仅通过一种方法获得结果^[83]。

A.1.2 校准标准的选择

一个重要的决定是,所有实验室是否宜使用相同的校准样品,或者实验室可以自由选择校准样品。使用单一校准样品减少了来自不同供应商的不同的校准样品引起的变异。另一方面,在这种单一校准样品中任何偏倚都会转化为标准值中相同偏倚。因此,使用单一校准品时需要非常仔细地对该校准样品进行定值,在许多情况下需要独立确认纯度或组成。允许实验室自由选择校准样品是在组织管理上更容易,不需要独立的确认试验,并允许实验室在没有改变的情况下应用自己的程序,但意味着在研究中宜包括某些核查校准适宜性的方法。作为一般准则:

- 对于成熟的测量,经验表明,现有校准标准的质量是有保证的,让实验室自由选择标准通常更可取;
- 在对市场上的标准质量存在重大怀疑的情况下,有必要确定使用统一的标准。

在某些情况下,不同的商业标准生产商从同一家公司获得其纯材料。因此,确定原始来源是否不同是有益的。

A.1.3 实验室的选择

宜根据证明的能力选择实验室。证明能力的合适证据可包括:

- 能力验证的结果;
- 独立 CRM 的结果(可能是作为质量控制样品与候选 CRM 一起分发);
- 方法确认数据;

- 全面可靠的不确定度评定；
- 之前参加其他 RM 研究中为相同被测量定值；和
- 对所讨论的被测量测定，第三方评估是否符合 ISO/IEC 17025^[52]或其他相关标准。

通过外部评估实验室的表现可能与其他有能力实验室的表现有所不同。因此，除了第三方符合 ISO/IEC 17025 评估证据外，还需谨慎获取更多有关表现的信息。

RMP 宜确保每个实验室实施的测量均符合 ISO/IEC 17025。特别是，宜满足 ISO/IEC 17025 中关于工作人员能力、设备校准和方法与结果批准的条款(正式发布使用[52])。

在缺乏独立评估的情况下，宜获得有关实验室质量体系范围的信息。

注 1：在满足相应标准要求方面的潜在差距有时可由 RM 生产者填补(例如，如果实验室没有存档系统，RM 生产者可以归档实验室的原始数据)。

注 2：为每台仪器设备获得校准证书在许多情况下是不切实际的。质量控制样品结果的一致性可以作为对所有相关输入因子进行充分校准的证明。

A.1.4 独立数据组的数量

参加实验室的数量不如独立数据组的数量重要。单个实验室可以提供多个数据组，这些数据组都是通过独立的测量程序获得的。接下来的讨论集中在数据组上，不论每个数据组是否由不同的实验室提供，有些实验室可提供多个数据组，或者所有数据组都由同一个实验室提供。

如果在单个实验室中进行测量，完全独立的结果是很难实现的。RM 生产者宜严格审查 A.1.1 中所述的所有关键步骤的变化，核查是否有足够的方法证明每一步都没有偏倚。这包括核查各种测量程序是否使用了相同的关键化学品、设备、校准样品等。组织者还宜在研究开始时提醒实验室不要筛选数据，即在各种程序之间的结果进行交叉核查之后，不要在没有通知的情况下删减/更改来自一个程序的数据。

RM 生产者宜设定一份记录在案、技术有效结果的最低数量，以便考虑赋值。数据组的数量宜足够大，以便考虑到报告失败的可能性、因技术原因剔除结果和预期统计评价之后，对估计值提供足够小的不确定度。

实验室间研究使用一个检测实验室网络，定值宜包括五个或更多的参加者提供技术上有效的数据。

注：定值不确定度小于实验室间再现性标准差的三分之一时至少需要 9 个参加者，除非实验室的能力卓越。

为达到标准值期望的不确定度，要考虑下列影响到所需数据组数量的因素。

- 所需的不确定度：赋值的不确定度通常随着技术上有效数据组的数量而减少，对于较小的不确定度需要更多的数据组。
- 技术上的问题：测量程序越不完善/成熟或技术上的挑战性越大，数据组之间变异的可能就越大。
- 技术上无效结果的可能性：即使是经验丰富的实验室也可能产生无法用于认定的技术上无效的结果。此类结果可能是由于不熟悉的材料、新的或修改的测量程序、挑战性的测量或不同寻常的报告要求。在技术错误风险较高的情况下，宜增加参加者或(单一实验室研究)的独立测量的数量。
- 再现性/重复性比率：如果已知数据组间变化是变化的主要来源，宜优先考虑更多的数据组，而不是对每个数据组进行更多的重复次数。
- 统计评价：不同的数据处理方法可能需要大量的观测，以提供良好的数值稳定性和/或获得足够小的不确定度。
- 通过 CRM 确认结果：由一种或多种类似 CRM 获得高质量的一致性结果，用于每个实验室的质量保证、证明所有数据组的有效性、并提供实验室和测量程序之间的偏倚信息。

A.1.5 单元数和重复测定次数

发送给每个参加者的候选 CRM 单元数,每个参加者执行重复测定次数以及这些测定的条件由实际以及评价考虑因素决定。

- 如果 RM 各个单元间的变异很大,那么对几个不同单元的单次测量比对单个单元几次重复测量更可取。如果污染、破损或不均匀性不是问题,送一个单元就足够。
- 在参加者缺乏可靠的不确定度评定的情况下,要求在中间精度条件下进行测量,可以提供参加者结果的可靠性指标或对参加者不确定度估计值的可靠性进行核查。
- 如果每个实验室收到一个以上的单元,为防止一个单元发生破损后,仍可以对其余单元进行较多的测量,从而不需要重新发样了。

A.1.6 质量控制样品

在质量控制中加入额外的样品是非常有益的。这些样品的结果可以识别技术问题,有助于技术评价。

- RM,特别是自然基体 RM 和质量控制样品,与尚未定值的材料一起测量时可以用来证明测量结果的有效性。
- 基质加标、空白加标等,可以用来核查测量程序的一部分、或者协助给材料赋值。
- 空白基体材料、空白提取物等,可用于证明当目标特性不存在时(如在成分测量中经常做的那样)测量程序提供的结果与零没有显著不同、或者建立校正或校正因子加上校正因子的不确定度。

注:有时用于实验室间研究质量控制的 CRM 不提供原始标签,以避免被识别。然而,由于可靠的 CRM 数量是有限的,有经验的实验室可能会从外观或测量值对材料加以识别。

A.1.7 给参加者的指导书

给参加者的指导宜包括:

- a) 研究目标的清晰概述;
- b) 不要与其他参加者比对结果,包括劝阻串通的理由(即合作交换信息);
- c) 要检测的单元数;
- d) 要执行的重复测定次数;
- e) 要使用的测量程序的任何限制或具体细节;例如,需要事先干燥和湿度校正;
- f) 最小取样量;
- g) 对测量结果的质量和溯源性的要求;
- h) 时间表(样品的分发、结果的报送);
- i) 发送方式;
- j) 样品中间贮存说明;
- k) 样品处理的具体说明,如适用;
- l) 关于质量控制措施的说明,以确定潜在偏倚;
- m) 生产者关于实验室识别和数据使用的政策信息,例如是否标明实验室身份、结果中是否将标明特定实验室以及结果是否用于定值研究以外的目的。

注 1:与参加讨论的实验室/组会议的(在分发样品和执行测量之前)可以帮助所有参加者协调在合作研究期间进行的所有行动,并讨论可能出现的问题或隐患。

在实验室间研究中,RM 生产者宜采取合理措施防止实验室之间的串通,包括但不限于上述 b)。

注 2:为每个参加者提供的材料使用不同的标签能使参加者更难比对结果。

A.1.8 报告

使用预先格式化的报告报表可能是有用的,因为它具有格式化报告的优点,且结果可以复制(如果以电子方式传输),从而减少了 RM 生产者在整理结果时的抄写错误。格式化报告形式的缺点是它们往往使实验室改变他们的报告形式,从而导致转录错误。如果以电子方式提交报告,则宜遵守 ISO 17034 关于电子记录的完整性的要求,特别是对检测结果报告的要求。

注:当报表以电子表格形式提交时,使用电子表格软件中的“锁定”或“保护”设置,通常可以防止无意的变更。

可以要求实验室报告各个结果(不仅是所有样品的平均值),无论是否报告不确定度陈述,即使报告平均值和扩展不确定度及其包含因子就足够了。

如果有校正已知程序偏倚的选项,如提取回收率, RM 生产者宜清楚地说明结果是否校正。RM 生产者也可能要求参加者报告偏倚核查(如通过添加),并使用这些来校正检测到偏倚的结果。如果参加者采用校正,任何报告的不确定度宜包括与校正相关的不确定度。

如果出现下述这样的结果,宜告知参加者如何报告接近检出限的结果。一方面,结果“小于”检出限会使统计评价更加困难。另一方面,接近检测限的报告结果与许多实验室的质量程序相矛盾。当结果有可能的在检出限附近时, RM 生产者宜要求实验室报出观测到的结果(包括负的),无论是代替他们正常的报告还是正常报告的基础上,或宜采用统计程序,允许“左删失”的结果,如“小于”声明,或仅允许结果大于零。

给参加者的指导书宜规定测量单位和报告定量结果的有效数字位数。

建议将所用测量程序大纲进行详细报告,以便了解测量过程所有阶段(例如,在化学分析中样品溶解/提取、目标分析物的分离、净化和定量)。适用时,宜要求参加者提供参考文献。

A.2 评价

A.2.1 评价的总体考虑

在评价过程中,可能会出现异常情况,需要与有关参加者进行沟通。生产者可与参加者联系,协助调查评价过程中任何阶段的异常。如果与参加者接触,建议初次接触不宜说明异常的性质(如结果偏离的方向);相反,宜首先邀请参加者调查并报告发现的任何错误。

如果由单个实验室提供来自多个测量程序的数据组,初始检查宜单独考虑各个数据组(即如同来自不同实验室的独立报告)。尽管如此,在得出发现异常时宜保留哪些数据组的结论时,标准样品生产者宜考虑某个实验室提交的所有数据组。

A.2.2 初步筛选

A.2.2.1 概述

每个参加者的信息宜在收到时或在收到后尽快进行审查。

对各参加者结果的初步审查宜核查基本报告或程序上的错误的证据,如遗漏的数据(包括任何要求信息的缺失)、不正确的重复次数、不适当的测量条件(如重复性与再现性条件)、检测项目的不正确标识(如通过偶然的错误标签)和测量单位的误报。在这个阶段,明显的错误宜在可能的情况下,迅速提交给参加者以进行检查和可能的纠正(见 A.2.1)。

在收到结果时发现明显高或偏低,可在本阶段交由参加者进行核查。

A.2.2.2 技术评价

识别潜在问题的进一步技术审查可包括(但不限于)通过所用技术(测量程序和原理、样品前处理方法等)或校准技术对结果进行分组。此外,将扩展的不确定度与提交结果平均值的置信区间进行比

较,说明陈述的不确定度是否符合实际,因为扩展不确定度至少宜与置信区间一样大。

对各参加者结果的技术审查,宜尽可能核查程序中错误的证据,如使用了不合适的测量程序或预处理程序、不适当校准、不适当的测量条件(如重复性与再现性条件)和不正确的测量单位。如果报告了测量不确定度,审查宜检查与典型期望相比的异常高或低的不确定度,以及在预期不确定度可获得的情况下,是否遗漏了重要的不确定度贡献量或用了不适当的不确定评定方法。

技术上无效的结果宜从数据组中剔除或纠正,如果可能且有必要的话,重新进行测量。

注1:技术上无效的结果不一定是离群值,也不是每一个离群值在技术上都是无效的。即使得到结果的条件明显不好,结果也可以很好地落在有效结果的范围之内。相反,与所有其他结果显著偏离的结果可能是数据组中唯一技术上有效的结果。

注2:即使报告中的不确定度显示出低估的证据,但只要赋值的统计程序选择恰当,仍可以进行有效赋值。参考文献[34]对此类程序提供进一步指导。

A.2.3 统计评价

A.2.3.1 统计评价的一般原则

通过协作研究或多个测量程序进行定值的目的将数据组间的偏倚随机化。统计评价通常假设被测量真值对应于一个总体参数的真值,通常是总平均值。

用来估计特定被测量的不同的程序,不论该被测量是否是操作定义的,可能存在系统偏差、也可能表现为每个数据组的实验室特定偏差。在评定测量不确定度时,通常宜考虑方法间差异的可能性。

如果报告了测量不确定度,统计评价宜核查异常高或低的测量不确定度,不确定度/位置异常,如与所报告的不确定度相比,远离任何中心估计值的结果,以及任何一个不确定度评定普遍不充分的证据(如比报告的不确定度更大的离散性)。与报告测量不确定度有关的异常情况宜尽可能加以解决,例如,可提供给参加者以进行核查和可能的校正。

A.2.3.2 分布

为期望值找到适当的估计值与(假设的或确定的)值所在的分布密切相关。

由数据组确定概率密度函数明显比实际获得数据需要更多的数据,然而根据定值研究中通常所获的数据组数量,有可能核查与假设分布的一致性并根据分布计算参数(如正态分布的平均值)。

因而,RM生产者宜核查是否有证据表明偏离假设的分布,例如,使用目视法(直方图、核密度图和正态概率图、或者更普遍的Q-Q图)或对偏离特定分布的统计检查,包括正态性检验或偏度和峰度的测定。对于远高于定量限的结果,经常可观测到数据组的近似正态分布;其他分布包括泊松分布(如微生物计数)或威布尔(Weibull)分布(如陶瓷的机械损伤)。所选定的分布宜与所报告的数据以及所涉测量的理论知识和历史知识一致。如果这些差异很大,除非给出意外分布的技术原因,否则不宜赋予任何值。

示例:理论上的探讨以及大量实验室间比对表明,土壤中微量元素的结果服从正态分布(除非接近定量限水平)。观测数据偏离正态分布(如拖尾)表明存在技术问题。如果不知道问题出在主要部分还是尾部,则不予赋值。

在某些情况下,结果可以转化为近似正态分布。一些常用的转换包括对数、平方根和指数形式。这种转变宜有技术基础,因为偏离期望分布的结果表明可能存在技术问题。

A.2.3.3 离群值

在协作研究的所有层面都可能产生离群结果:单次观测、分组观测(如按瓶分组)、或从整个方法/实验室的结果都可以观测到是离群的。可以通过以下方法识别离群值:对离群数据的均值和方差进行适当的离群值检验、原始数据的图形检查、以及使用曼德尔的 h 和 k 统计量^[24]。

离群观测值或平均值不宜仅仅基于统计离群值检验而剔除,但如果技术原因支持,则可以剔除。

注1: 技术原因包括不适当校准、不合适的测量程序、试剂使用不当、未能考虑干扰、偏离独立质量控制样品的标准值等。

注2: 离群值检验通常不考虑测量不确定度。当考虑到各自的不确定度时,离群数据组可以与其他数据相一致。

注3: 不确定度异常大的数据点,如果在不确定度范围内仍与其他结果一致,并且如果与其他报告的不确定度相比,不确定度太大,表明存在技术问题,则可以基于技术原因而剔除。它们也可以保留,因为它们与其他结果一致。

注4: 显示高离群方差的数据组可能表明方法重复性或中间精密度欠佳,可以技术理由予以剔除。

注5: 不同测量程序的精密度可能有差异,为了保证程序的代表性,保留相对不精密程序的数据组也很重要。

注6: 极端方差的特殊情况是一组具有零方差的结果。当实验室报告的有效位数太少时可能出现这种情况。这可以通过给参加者适当指导(A.1.7)来防止、或通过参照参加者来修正。使用这样一组相同数据的平均值或中位数也是有效的,例如,当组间影响决定使用组均值时。

A.2.3.4 稳健统计

稳健统计提供了大量的统计方法,这些方法明确允许在近似正态分布的数据组中存在离群值。通常,稳健法分配的权重随着偏离数据主体的距离而减小,它常用于估计简单的被离群值影响的数据组的均值和标准差,以及许多其他感兴趣的参数。其他情况下也存在稳健方法,例如,具有明显不同不确定度的多组报告值、或者对每个数据组由多个重复测量数据组成的数据分析。

稳健估计量对极端离群值的影响具有很强的抵抗力。对于对称的“重尾”分布(包含更大比例远离正态分布预期均值的值),稳健统计通常提供了比简单算术平均值具有更小方差的均值无偏估计。这种方法的结果是在同样的情况下会给出比算术平均值更小不确定度的无偏估计。对于具有极端离群值的数据,稳健统计比平均值具有很大的改进。在B.5中给出了有用的稳健统计方法的简短介绍,以及其在RM定值中的使用条件和对更详细描述参考文献。

注: 稳健统计对极端值的容忍也有助于在包含若干极端值的较大数据组中识别离群值。

A.2.3.5 分组(“聚类”)

统计评价宜核查结果分组的发生情况,如按测量程序、校准样品、试剂或地域等。

- a) 如果不同组的均值之间的差异在统计上是显著的,并且因差异过大而不能允许对材料的预期用途有足够小的不确定度,那么就不能赋予单一的特性值。如果沿试剂/校准样品分组,技术评价宜核查是否所有的试剂/校准样品都合适。如果分组是按测量程序进行的,生产者可以为每个测量程序提供赋值。
- b) 如果存在显著差异,这些组的均值之间的差异相对较小,则可以赋予单一值。然后,宜在定值不确定度中添加一个说明组间的变化的额外不确定度项。文献中描述了解决该估计问题的几种方法^[53-54]。
- c) 如果这些群之间的差异很大,并且这些群与测量程序或对差异的其他技术解释亦不相关,则不能赋予任何值。为了克服现有测量程序较差的一致性,可能需要更多的结果。

注: 可见的分组可能来自其他原因,包括偶然的(特别是小数据组)、应用程序的区域差异或不同校准标准的使用。

A.2.4 赋值(加权/非加权均值)

当数据组均值服从近似正态分布且不采用加权时,数据组均值的非加权算术均值作为所赋予的值 y_{char} 。

$$y_{\text{char}} = \frac{\sum y_i}{p} \quad \dots\dots\dots (\text{A.1})$$

式中:

y_i ——第 i 个数据组的均值;

p ——数据组数。

如果数据组均值之间的差异与每个数据组内的变化的影响相比不显著,各个结果的均值也可以用于赋值。

加权均值通常用公式(A.2)的一般形式进行计算:

$$y_{\text{char}} = \frac{\sum_{i=1}^p w_i x_i}{\sum_{i=1}^p w_i} \dots\dots\dots (\text{A.2})$$

式中, w_i 是应用于每个数据组均值 x_i 的权重(或者给单个值提供值和不确定度)。

最简单权重 w_i 的选择见公式(A.3):

$$w_i = \frac{1}{u_i^2} \dots\dots\dots (\text{A.3})$$

式中, u_i 是为值 x_i 报告的标准不确定度。只有当报告的不确定度被证明是可靠的,并且清楚地表明报告的不确定度可以说明组间差异时,才宜使用这种方案。不确定度不能充分考虑观测到的值 x_i 离散时,宜使用另一套方案。例如,在参考文献[33]和[34]中描述了不易受到不可靠的不确定评定影响的替代方法。如果不能赋予可靠的权重,简单均值通常是一个有用的保守替代方案。

如果结果不服从正态分布但能将其转化成正态分布的数据,则可以按以下步骤对数据进行评价:转化原始数据;运用公式(A.1)进行计算;计算赋值和置信区间;通过逆向变换成赋值和置信区间。如果还需要标准不确定度,可以参考分布方差的统计文本或者使用 GUM 来获得它,要记住对于超过 15% 和更高的相对不确定度,一阶误差传播可能会严重失败,并且可能需要更高阶项。

如果结果不服从正态分布,且不能转换成正态分布的数据,则宜采用与观测分布一致的统计上合理的方法。

A.2.5 不确定度

A.2.5.1 使用方差分析进行不确定度评定

方差分析(ANOVA)可以用作处理数据的工具。在评估不确定度分量,如瓶间均匀性或实验室间标准差时,使用方差分析特别有用。否则,作为替代策略,只能计算均值的均值。

A.2.5.2 基于不确定度的评定

理论上可以将结果,包括它们的不确定度,合并成一个单一值(特性值)和一个合成标准不确定度。

所描述的方法包括由不确定度加权的结果^[35]、确定不确定度的详细表达式^[34]、最小二乘法拟合(如参考文献[36])或将不确定度分解为公共部分和个别部分^[37]。

注:这些方法在理论上是可能的,但在实践中往往很难实现,迄今为止很少使用。

A.2.5.3 实验室未提供不确定度的评定

如果用公式(A.1)计算标准值,数据组均值近似服从正态分布,且不采用加权,那么, p 个数据组均值的均值标准差可用作 u_{char} 。

$$u_{\text{char}} = \frac{s(y)}{\sqrt{p}} = \frac{1}{\sqrt{p}} \cdot \sqrt{\frac{\sum (y_i - y_{\text{char}})^2}{p-2}} \dots\dots\dots (\text{A.4})$$

式中, $s(y)$ 表示 p 个数据组均值的标准差。

如果结果不服从正态分布并不能转化为正态分布的数据,采用与观测到的分布一致的计量学上和统计上合理的方法。

注:公式(A.4)只覆盖实验室间部分的不确定度。如果许多或所有的数据组是通过相同材料校准得到的,校准样品赋值的不确定度与由公式(15)得到的结果相比是显著的,则要增加校准样品赋值的不确定度。

A.3 多用途协作研究的使用

A.3.1 实验室间比对可能有不同的目的,包括:标准样品赋值、实验室能力评价和测量程序性能特征评价。一般来说,一个特定的研究最好只适用于其中的一个目的;否则,可能为了有利于一个目标而损害了另一个目标,或者混淆了参加者研究的目的。尽管如此,考虑将 RM 的特性研究与其他研究结合起来有利于节省成本,只要适当注意,避免主要缺点和满足某些条件。A.3.2 提供了主要缺点的指南;A.3.3 提供了这种研究与 RM 定值进行组合的条件。

A.3.2 将 PT 或方法性能研究与定值研究合并,潜在的不兼容或矛盾包括以下方面。

- a) PT 研究的目的是评估实验室的能力,而定值研究的参加者仅限于证明具有能力的实验室。因此,将 PT 研究与定值研究相结合,是假设为定值目的证明能力的事实宜放在 PT 中评估。
- b) 能力验证通常不要求进行大量的重复测量,也很少规定所有参加者进行等同的重复测量。
- c) PT 结果通常被视为机密,无法保证获得实验方法的详细信息。
- d) 参加者通常支付参加 PT 研究的费用,参加的动机有时可以由商业驱动。因此,参加者通常不愿意遵守定值研究中所要求的详细研究安排、对溯源性和质量保证的要求、校准、提供不确定度等。
- e) PT 的结果可用于向潜在客户、认可机构和其他方面证明能力。因此,参加者很少愿意自由讨论,并承认将定值研究中所要求的技术问题对结果进行技术评定。
- f) PT 研究通常通过与其他实验室或外部设定的标准相比评估实验室的能力,而定值研究的目的是获得对真实及其不确定度的良好估计。这导致采用不同的评价方法,特别是在处理极端值/离群值时。
- g) PT 研究通常对给愿意参加的任何实验室是开放的。因此,RM 生产者通常难以控制参加者应用的测量程序,即使存在其他程序,最终也可能只收到一种程序的结果。
- h) 在 PT 研究结束之前,无法开始反馈那些为了检查异常的结果,而在定值研究中,要求跟踪明显的异常结果,要求随时反馈信息。
- i) 对 PT 研究的评价也非常注重统计,而定值研究则更侧重于技术评价。
- j) 方法性能研究在开发之后立即评估测量程序的性能,因此很少有关于典型性能的先验信息。
- k) 方法性能研究通常只适用于某个测量程序,因此不适合用于不依赖方法的被测量的赋值。

A.3.3 由于这些潜在的矛盾,如果标准样品使用从 PT 计划参加者处收集的数据进行认证,RMP 宜:

- a) 在研究开始前,决定哪些指定实验室的数据组将用于赋值;
- b) 证明这些实验室具有独立于本研究的能力,例如,通过前几轮计划的表现或其他方式;
- c) 按照 A.1.3 列出的标准,组织这些选定的实验室进行研究;
- d) 按照 A.2 列出的标准,评定这些实验室的结果;
- e) 通知被提议的参加者,结果将用于认证研究,并取得获取技术细节的许可。

附录 B

(资料性)

统计方法

B.1 单因子方差分析(ANOVA)

假设有 a 个组,每个组有 n_i 个成员。理论上各组成员数量宜相等,但实际上情况并不总是这样,一些数据由于各种原因会产生缺失。存在缺失数据时的计算方法见[55]和[56],并建议用来处理不完整数据组。然而数据组越不完整(缺失数据的数量越多),估计的质量就会越差。

应用于标准样品研究的单因子方差分析通常基于以下统计模型:

$$x_{ij} = \mu + \delta_i + \epsilon_{ij} \quad \text{.....(B.1)}$$

式中:

x_{ij} ——第 i 组第 j 次观测值;

μ ——总体可能结果的真均值(假设由此产生观测到的结果 x_{ij});

δ_i ——第 i 组均值与 μ 的(真)偏差;

ϵ_{ij} ——第 i 组中第 j 次观测的随机误差项。

假设 δ_i 来自于均值为 0、标准差为 σ_{between} 的总体(通常服从正态分布), ϵ_{ij} 来自于均值为 0、标准差为 σ_{within} 的总体(通常也服从正态分布)。需要注意的是,对于所有组,假设每组的真标准差都是相同的。这两种标准差说明了观测数据的离散程度,这里应用单因子方差分析的目的在于估计这两种标准差。

数据的离散程度可以用平方和的和来表示,也可称为“离均差平方和”。这些离均差平方和表示了方差分析中不同等级的离散程度[55]。由电子表格程序或者统计软件得出组间均方(M_{between})与组内均方(M_{within})可以按公式(B.2)至公式(B.4)转变成下列估计的方差:

$$s_{\text{within}}^2 = M_{\text{within}} \quad \text{.....(B.2)}$$

$$s_{\text{between}}^2 = \frac{M_{\text{between}} - M_{\text{within}}}{n_0} \quad \text{.....(B.3)}$$

其中

$$n_0 = \frac{1}{a-1} \left[\sum_{i=1}^a n_i - \frac{\sum_{i=1}^a n_i^2}{\sum_{i=1}^a n_i} \right] \quad \text{.....(B.4)}$$

s_{between} 和 s_{within} 分别是对 σ_{between} 和 σ_{within} 的估计。

在每组包含 n 次观测的研究中,如果没有缺失数据存在时, n_0 就等于 n 。

如果有必要,也可根据公式(B.5)估计总平均值 $\bar{x}$:

$$\bar{x} = \frac{1}{\sum_{i=1}^a n_i} \sum_{i=1}^a \sum_{j=1}^{n_i} x_{ij} \quad \text{.....(B.5)}$$

当组间效应不存在时, s_{between} 预计为(接近)零。如果由于实验的原因 s_{between}^2 得到了负值,那么宜设置为零。

注:当 s_{between}^2 的初始估计值为负数时,那么 $s_{\text{between}} = 0$, $s_{\text{within}} = s(x)$ 。其中 $s(x)$ 是所有观测值 x_{ij} 的标准差,也是两个参数的极大似然估计值(等同于限制极大似然估计值),这同样是统计上有效的程序。

B.2 嵌套设计的双因子方差分析

当协作研究的结果被用于确认材料的均匀性以及对其定值,并且各组间重复性恒定时可以使用这

种模型。对于实验室间研究这一特定情况,试验方案说明见图 B.1。当活动由不同的测量程序组成时,活动的布局基本相同。

结果可用模型公式(B.6)表示:

$$x_{ijk} = \mu + A_i + B_{ij} + \epsilon_{ijk} \quad \dots\dots\dots (B.6)$$

式中:

x_{ijk} ——由程序/实验室 i 报告的第 j 个样品单元的第 k 个结果;

μ ——总体可能结果的真均值(假设由此产生观测到的结果 x_{ijk});

A_i ——程序/实验室 i 的误差;

B_{ij} ——在程序/实验室 i 内的第 j 个样品单元的误差;

ϵ_{ijk} ——在 ij 组的第 k 次观测的随机误差项。

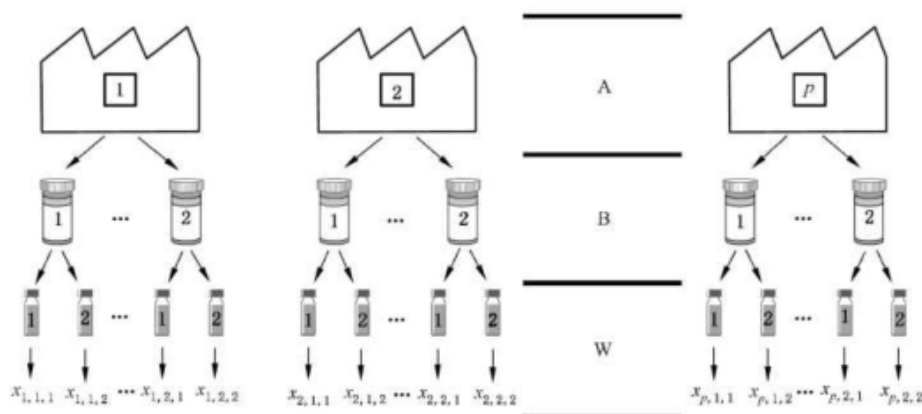
假设 A_i 来自于均值为 0、标准差为 σ_L 的总体,假设 B_{ij} 来自于均值为 0、标准差为 σ_{lb} 的总体,假设 ϵ_{ijk} 来自于均值为 0、标准差为 σ_r 的总体。如前所述,假设测量误差项不仅在每个 RM 单元内、而且在每个实验室内都具有相同的方差。通常假设总体服从正态分布。需要估计的参数为实验室间标准差 s_L 、组间标准差 s_{lb} 和重复性标准差 s_r 。公式(B.7)显示了它们与误差项的关系。

$$\begin{aligned} s_L &= \sqrt{\text{Var}(A_i)} \\ s_{lb} &= \sqrt{\text{Var}(B_{ij})} \\ s_r &= \sqrt{\text{Var}(\epsilon_{ijk})} \quad \dots\dots\dots (B.7) \end{aligned}$$

对于单元间均匀性的 s_{lb} ,与均匀性研究本身一样(见 7.8),同样也要考虑不能检出批不均匀性的情形。

当有足够多的相同重复次数的试验结果(每个单元有相同数量的重复测定,每个程序/实验室具有的单元数也相同),技术上和统计上无效的结果被剔除之后,所有这些参数可以采用方差分析方法同时估计^[55]。如果单因子方差分析的要求由于无效和丢失的结果无法满足,那么单元间的方差的显著性需要用其他方法确定(见第 7 章)。

关于平衡和非平衡方差分析的理论细节和其他方法见标准教科书(见参考文献[15]、[57]、[58]和[59]),有关标准样品定值中方差分析的讨论已在参考文献(见参考文献[16]、[55]、[60]和[61])中讨论。



标引符号说明:

A ——实验室间变异;

B ——单元间变异(图中大容器表示 RM 单元);

W ——测量重复性(图中小容器表示单个提取物)。

图 B.1 结合批均匀性研究的协同试验示意图(RM 的定值双因子设计图)

计算上述估计值 s_L 、 s_{lb} 和 s_r 的公式^[13,16,62]如下,总体平均值 $\bar{x}$ 可以使用公式(B.8)计算。

$$\bar{x} = \frac{1}{\sum_{i=1}^p \sum_{j=1}^{b_i} n_{ij}} \sum_{i=1}^p \sum_{j=1}^{b_i} \sum_{k=1}^{n_{ij}} x_{ijk} \quad \dots\dots\dots (\text{B.8})$$

式中:

p ——实验室数;

b_i ——第 i 个程序/实验室使用的单元数;

n_{ij} ——单元 ij 的重复测量次数。

s_r 、 s_{bb} 和 s_L 可由公式(B.9)至公式(B.11)得到:

$$s_r^2 = M_{\text{within}} \quad \dots\dots\dots (\text{B.9})$$

$$s_{bb}^2 = \frac{M_B - M_{\text{within}}}{n_0} \quad \dots\dots\dots (\text{B.10})$$

$$s_L^2 = \frac{M_L - n_0' s_{bb}^2 - M_{\text{within}}}{(nb)_0} \quad \dots\dots\dots (\text{B.11})$$

其中

$$n_0' = \frac{\sum_{i=1}^p \left(\frac{\sum_{j=1}^{b_i} n_{ij}^2}{\sum_{j=1}^{b_i} n_{ij}} \right) - \frac{\sum_{i=1}^p \sum_{j=1}^{b_i} n_{ij}^2}{\sum_{i=1}^p \sum_{j=1}^{b_i} n_{ij}}}{p - 1}$$

$$n_0 = \frac{\sum_{i=1}^p \sum_{j=1}^{b_i} n_{ij} - \sum_{i=1}^p \left(\frac{\sum_{j=1}^{b_i} n_{ij}^2}{\sum_{j=1}^{b_i} n_{ij}} \right)}{\sum_{j=1}^p b_i - p}$$

和

$$(nb)_0 = \frac{\sum_{i=1}^p \sum_{j=1}^{b_i} n_{ij} - \frac{\sum_{i=1}^p \left(\sum_{j=1}^{b_i} n_{ij} \right)^2}{\sum_{i=1}^p \sum_{j=1}^{b_i} n_{ij}}}{p - 1}$$

均方(M)宜采用已发布的方法(如参考文献[16]和[62])或者合适的统计软件包获得,所给的表达式用于说明计算遗漏和/或剔除(无效)数据。对于完整的数据组,可以使用 ISO 5725-3^[14]中较简单的公式。

在这里计算出来的 s_{bb} 可以被用来估计由不均匀性引入的不确定度,计算与第 7 章相同。

当应用于实验室间研究时,核查组内恒定方差的假设是很重要的。该假设可以采用统计检验进行核查,如巴特利(Bartlett's)检验或者莱文(Levene's)检验。

注:相同的模型可用于处理图 5b 所示形式的单元内均匀性研究中的数据,即用 RM 单元内等分试样替换上述的 RM 单元。在研究单元内均匀性时,公式(B.11)估计单元间 s_{bb} ,公式(B.10)估计单元内 s_{wb} (其可用于标准值不确定度中分量 u_{wb}),公式(B.9)估计测量重复性标准差。

B.3 线性回归(单变量线性模型)

B.3.1 基本模型

标准样品研究中使用线性回归方法确定简单的变化速率,并检验其统计显著性。本节描述最简单

的情况。

简单线性模型可用公式(B.12)表示:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{.....(B.12)}$$

式中, β_0 是截距, β_1 是斜率。 ε 表示的是随机误差项, 通常假设服从均值为零的正态分布。

给定的一组 n 对成对的观测值, X 和 Y , 各个观测值 (x_i, y_i) 之间的关系如公式(B.13)表示:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{.....(B.13)}$$

示例: 对于一个基本的稳定性研究, x_i 表示的是观测的时间, y_i 表示感兴趣特征的观测值。

B.3.2 拟合模型

回归的参数可以由公式(B.14)和公式(B.15)计算:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{.....(B.14)}$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad \text{.....(B.15)}$$

式中, b_0 和 b_1 分别是截距和斜率的估计, $\bar{x}$ 和 $\bar{y}$ 分别是相应观测值的均值。

b_1 和 b_0 的标准误差 $s(b_1)$ 和 $s(b_0)$ 可以利用公式(B.16)和公式(B.18)计算:

$$s(b_1) = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad \text{.....(B.16)}$$

其中

$$s^2 = \frac{\sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2}{n - 2} \quad \text{.....(B.17)}$$

同时

$$s(b_0) = s(b_1) \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}} \quad \text{.....(B.18)}$$

B.3.3 假设检验和核查

初步拟合模型的典型审查包括:

- 计算残差 $[r_i = y_i - (b_0 + b_1 x_i)]$, 并做残差与时间的图, 然后检查曲率、运行间效应、各时间点数据离散程度的显著性检验或者离群值的检验;
- 绘制残差的正态图(或 Q-Q 图), 并检验非正态性的证据;
- 或者进行组间的离散程度显著性检验、组间残差的显著性检验、线性检验。

B.3.4 显著性变化的统计检验

对于以上简单情况, 在确认了基本的回归假设的有效性之后, 通常用 t 检验去检验斜率是否是显著不同于零。这可以通过计算公式(B.19)中的 t 统计量进行:

$$t_{b_1} = \frac{|b_1|}{s(b_1)} \quad \text{.....(B.19)}$$

然后将该统计量与 t 分布当自由度为 $n-2$, 95% 置信水平的双尾临界值做比较。如果计算的检验统计量 t_{b_1} 超过了临界值, 那么斜率就被认为在 95% 置信水平下显著不同于 0。

B.3.5 回归直线的置信区间

拟合值 b_0 和 b_1 可用于提供对某个特定值 $\hat{x}$ 的响应值的估计值 $\hat{y}$, 利用:

$$\hat{y} = b_0 + b_1 \hat{x} \quad \dots\dots\dots (\text{B.20})$$

有时候获得一个预测值 $\hat{y}$ 的置信区间是有用的。在 95% 置信水平双侧置信区间可以表示为:

$$\hat{y} \pm t_{95, n-2} s \sqrt{\frac{1}{n} + \frac{(\hat{x} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad \dots\dots\dots (\text{B.21})$$

式中, $t_{95, n-2}$ 是自由度为 $n-2$ 的双尾 t 分布的临界值。不同置信水平的置信区间可类似于用合适的 t 分布临界值所获得。

注: 该区间并不是对个体值 y_i 的预测区间。

B.4 根据稳定性研究结果预测保质期或者选择初始监测点

B.4.1 原理

如果储存期间特性值随时间的变化有一个可接受的范围(由规定的上限和/或下限所描述), 则使用稳定性研究结果可能有助于估计标准样品的贮存寿命(或“保质期”)或者设置初始监测点。

其原理是考虑到估计的降解率, 估计未来值 95% 置信区间, 并选择其中一个置信限(上限或下限)与规定限相交的最短时间(设为 RM 的贮存期), 该原理如图 B.2 所示。

针对某个 CRM, 选择标准值变化的任何可接受范围宜小于扩展不确定度, 理想情况下宜小于 $u_{\text{CRM}}/3$ 。对其他标准样品, 任何可接受的变化宜足够小, 以避免对预期用途产生不利影响。

B.4.2 线性趋势下保质期预测

设 RM 特性值可接受范围的上限和下限分别为 L_{upr} 与 L_{lwr} , 当降解按照简单线性形式进行时, 预测 RM 未来值的置信区间由公式(B.20)和(B.21)给出, 以 x 作为稳定性研究开始的时间。通过将公式(B.21)设置到每个规格限值上并求解 $\hat{x}$, 以获得最短的正的保质期 t_{shelf} , 从而得到所需的保质期。也就是说, t_{shelf} 的选择是根据公式(B.22):

$$t_{\text{shelf}} = \min(t_{s, \text{upr}}, t_{s, \text{lwr}}) \quad \dots\dots\dots (\text{B.22})$$

式中, $t_{s, \text{upr}}$ 和 $t_{s, \text{lwr}}$ 是下列方程的最小正数解:

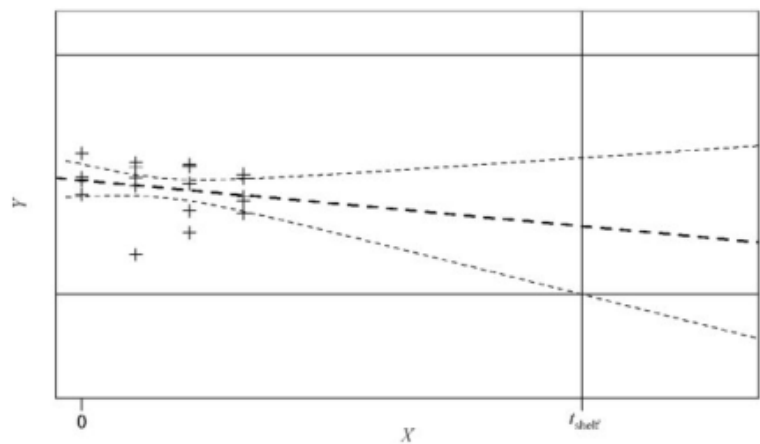
$$L_{\text{upr}} - \left[b_0 + b_1 t_{s, \text{upr}} \pm t_{95, n-2} s \sqrt{\frac{1}{n} + \frac{(t_{s, \text{upr}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right] = 0 \quad \dots\dots\dots (\text{B.23})$$

$$L_{\text{lwr}} - \left[b_0 + b_1 t_{s, \text{lwr}} \pm t_{95, n-2} s \sqrt{\frac{1}{n} + \frac{(t_{s, \text{lwr}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right] = 0 \quad \dots\dots\dots (\text{B.24})$$

注 1: 公式(B.23)和(B.24)一共会有四个实数解。当其中一个公式并没有得到正数解的时候, 例如统计上显著性降解会使得(B.24)产生两个正数解, 则忽略负解。

注 2: 公式(B.23)和(B.24)可以用代数或用迭代求解。

注 3: 如果规定限值太接近[特别是任何一个规定限公式(B.21)给出的区间里], 则不可能得到一个有价值的解决方案。



标引符号说明：
X ——时间；
Y ——值(任意单位)。
注：“+”表示在时间零点处稳定性研究的数据点。水平的实线表示规定的上限和下限。直虚线表示用 B.3 程序拟合的最小二乘法线。弯曲的虚线表示用公式(B.21)计算得到的 90%置信区间上限和下限。垂直实线位于保质期的时间(t_{shelf})[见公式(B.22)]。

图 B.2 根据稳定性研究计算保存期的原理

B.4.3 其他降解模型下保质期预测

对于不同的降解模型,保质期可以按照 B.4.2 所描述的同样的通用原理进行预测,并进行置信区间计算。公式(B.23)和公式(B.24)因此需要调整去适应特定的降解模型。

B.4.4 选择初始监测点

同样的原理可以用来选择发证后稳定性监测的初始监测点。选择特性值可接受的变化,以上计算出来的 t_{shelf} 就表示了从稳定性研究开始第一次要进行监测的时间。

B.5 稳健统计

B.5.1 实验室间研究的稳健估计量

表 B.1 中给出了一系列适用于实验室间研究的基本稳健估计量,以及文献来源和一些关键性能指标。这些估计量的使用需要服从以下条件:

- a) 每一个实验室数据宜由单一值表示(其本身可能是均值、中位数或者其他关于位置的估计值);
- b) 宜仔细检查来自实验室的信息与结果是否存在技术上的错误,技术上无效的结果在使用稳健方法之前需要从数据组中剔除;
- c) 在因技术原因剔除任何结果之后,通常数据点不宜少于 10 个;
- d) 数据组中大部分值(一般是中心部分)宜服从近似正态分布或者至少是对称分布(如果需要可转换后);
- e) 没有因测量程序或地理来源等引起的明显分组现象;
- f) 数据组中极端离群值的比例不宜超过所用估计量的崩溃点的一半。

表 B.1 中的稳健估计量并不是唯一的列表。算法 A 和汉佩尔(Hampel)估计量都可以通过改变调节参数进行调整,同时还存在很多可以替代的算法。算法上的差异显示,其他的稳健估计量由于被证明有相似或者比表 B.1 估计量更好的性能也可以被使用。其他估计量的性能可以用理论崩溃点、相对效率和适用于应用的其他附加特征(如抗不对称性或者对较小数据组的表现)进行评估。

如果 RM 生产者能够证明所选估计量在应用的典型数据组上的表现合适,则上述 c) 处的数据点数量可以小于 10。

注 1: 极端数值点的比例上限远低于估计量的崩溃点,不仅可以避免离群值的比例增长所带来的不利影响,同时如果这个比例过大,我们通常宜思考是否因为缺少实验室之间的一致性而终止认证活动。

注 2: 表 B.1 中的估计量并没有利用报告的不确定度或实验室内精密度的信息。因此,在实验室均值的不确定度不可得到或者不可靠时,这些估计量和最简单的算术均值是最合适的。

注 3: 很多稳健估计量可通过调整调节参数以降低相对效率为代价去增加对离群值的抗性。在效率(影响不确定度)和离群值的抗性之间最佳平衡取决于数据的分布,其中包括离群值。通过调节参数对正态分布给予相对效率为 0.8~0.95 是比较典型的。

表 B.1 一些稳健统计量及其特性

稳健方法	被估计的总体参数 (注 1)	相对效率 e_R (注 2,注 3)	崩溃点 (注 4,注 5)	参考文献 (注 6)
样本均值	均值	1.0	0%	
中位数	均值	0.67	50%	
算法 A	均值和标准差	$0.97^{a,b}, 0.73^c$	25%	ISO 5725-5 ISO 13528
汉佩尔(带 Q_n 或 $MADe$ 尺度)	均值	0.96^d	50%	ISO 13528
样本标准差	标准差	1.0	0%	
标准化四分位距 (nIQR)	标准差	0.37	25%	ISO 13528
中位绝对离差 (MADe)	标准差	0.37	50%	ISO 13528
Q_n	标准差	0.81	50%	ISO 13528
注 1: 严格来说,这里的总体服从正态分布且假设已经被离群值所污染。 注 2: 相对效率是对估计量用于正态分布数据时可变性的一个度量。它是估计量的方差除以相同数据的样本标准差的算术平均值或(对于标准差的估计量)的方差。效率越高越好。表中显示的效率是大样本效率。对于小样本,表中的估计量的相对效率相比于均值或者标准差略有提高。 注 3: 尽管对于正态分布数据的稳健估计量并没有均值那样有效(相对效率小于 1),但对于其他严重拖尾或者存在离群值的分布,稳健估计量的表现会迅速提升且超过均值。同样的道理也适用于样本标准差和相应的稳健标准差。 注 4: 崩溃点是估计量对某一比例极端离群值抵抗性的度量,值越高越好。 注 5: 这里用的崩溃点的定义是:当一个大的正态分布的数据组中一定比例数据点移向正无穷大,而估计量并没有也移向正无穷大,这个比例就叫作崩溃点。例如,如果正好有低于 50% 的数据被正无穷大值取代,那么中位数仍将保留在余下的有限值数据中。 注 6: 除了那些已经被熟知的均值、中位数和样本标准差之外,还提供了 ISO 标准的参考文献。 注 7: 表中的性能信息来自 ISO 13528:2015^[10]。 ^a 算法 A 的调整参数为 1.5,与 ISO 5725-5 和 ISO 13528 相同; ^b 算法 A 在调整参数 $k=1.5$ 时,均值估计值的效率与 ISO 5725-3 相同; ^c 算法 A 在标准差估计值的效率时,调整参数 $k=1.5$; ^d 汉佩尔算法的效率用调整参数 $a=1.5, b=3.0, c=4.0$ 计算出来的,与 ISO 13528 相同。				

B.5.2 稳健估计量不确定度的评价

从实验室数据直接得出的位置估计量的不确定度一部分取决于用什么估计量,另一部分取决于数据是来自于何种分布。对于正态分布数据,稳健估计量的不确定度为:

$$\frac{s^*}{\sqrt{e_R N}} \dots\dots\dots (B.25)$$

然而稳健估计量很少用于正态分布的数据,它们用于包含很大比例偏离于(真的)均值或者包含一些非常极端离群值的数据。在这些情况下,不确定度经常要高于公式(B.25)给出的值。最合适的不确定度很大程度取决于分布。从明显非正态分布数据评估估计值不确定度最普遍采用的方法是用模拟或者再抽样法。对于一个特定数据缺少由模拟或者再抽样法得到的不确定度时,稳健估计量的定值不确定度不宜小于公式(B.26)给出的值:

$$1.253 \frac{s^*}{\sqrt{N}} \dots\dots\dots (B.26)$$

公式(B.26)中因子 1.253 是依据在正态分布的大量结果中的中位值的标准差,或者中位值作为均值估计值的效率所得到的。然而更加精细的稳健方法可以得到比中位值更高的效率,这个因子(指的中位数)被认为是一个保守偏高的估计,它适用于适度的非正态性。

附 录 C
(资料性)
实 例

C.1 均匀性研究

在制备土壤中铬的有证标准样品项目中进行了单元间均匀性研究,其数据列于表 C.1。

表 C.1 土壤中铬单元间均匀性研究的测量数据

瓶号	结果 No.1	结果 No.2	结果 No.3
1	121.30	128.74	119.91
2	120.87	121.32	119.24
3	122.44	122.96	123.45
4	117.60	119.66	118.96
5	110.65	112.34	110.29
6	117.29	120.79	121.42
7	115.27	121.45	117.48
8	118.96	123.78	123.29
9	118.67	116.67	114.58
10	126.24	123.51	126.20
11	128.65	122.02	121.93
12	126.84	124.72	123.14
13	122.61	128.48	126.20
14	118.95	123.82	118.11
15	118.74	118.23	117.38
16	119.74	121.78	121.01
17	121.21	123.28	116.38
18	129.30	124.10	122.02
19	136.81	129.80	128.47
20	127.81	117.66	122.90

注：所有结果的单位为毫克/千克。

表 C.2 所示为利用诸如电子表格软件对表 C.1 的数据进行方差分析计算的结果。

表 C.2 土壤中铬单元间均匀性研究的方差分析表

变异来源	平方和(SS)	自由度(ν)	均方(MS)
单元间	1 037.1	19	54.59
单元内	330.5	40	8.26
总和	1 367.6	59	

组间方差(均匀性研究中的单元间方差)用公式(C.1)估算：

$$s_{\text{between}}^2 = \frac{M_{\text{between}} - M_{\text{within}}}{n_0} = \frac{54.59 - 8.26}{3} = 15.44 \text{ mg}^2/\text{kg}^2 \quad \text{.....(C.1)}$$

单元间标准差是该方差的平方根如公式(C.2)所示:

$$s_{\text{bb}} = \sqrt{15.44} = 3.93 \text{ mg/kg} \quad \text{.....(C.2)}$$

重复性标准差可由 M_{within} 计算得到,如公式(C.3)所示:

$$s_r = \sqrt{M_{\text{within}}} = \sqrt{8.26} = 2.87 \text{ mg/kg} \quad \text{.....(C.3)}$$

C.2 应用重复性欠佳的测量程序研究单元间均匀性

实例取自参考文献[63],一个有关猪肾组织 γ -谷氨酰转移酶催化浓度的认定。猪肾组织用氯化钠稀释并均匀化。重新沉淀后,产物依次经过 DEAE 纤维素层析柱(DEAE Trisacryl)和羟基磷灰石层析柱进行纯化。选择质量浓度为 60 g/L 的牛血清白蛋白作基体,因为它的加入并不改变部分纯化酶的催化活性。通过称量 101 个定期采集的安瓿来检查灌装程序。在灌装期间未发现趋势。最后冻干材料。

对材料的污染酶、纯度、含水量和残余氧含量进行核查,并查明材料适于作 CRM。

为量化单元间变异,评估均匀性的数据取自早先的定值演练。20 个安瓿每个测量 6 次。数据按安瓿分组,并对其进行方差分析。计算单元间标准差(s_{bb})、单元内标准差(s_r)。方差分析结果包括均方和不同水平的标准差估计值,如表 C.3 所示。

表 C.3 猪肾组织中 γ -谷氨酰转移酶催化浓度测量结果(括弧内是相对值)

平均值	67.78 IU/L
M_{between}	1.76 IU ² /L ²
M_{within}	1.63 IU ² /L ²
s_r	1.28 IU/L(1.88%)
s_{bb}	0.147 IU/L(0.22%)
u'_{bb}	0.196 IU/L(0.29%)

单元间标准差用公式(C.4)计算:

$$s_{\text{bb}}^2 = \max\left(\frac{M_{\text{between}} - M_{\text{within}}}{n_0}, 0\right) = \frac{1.76 - 1.63}{6} = 0.0217 \quad \text{.....(C.4)}$$

从而得到 s_{bb} 的估计值为 0.147 IU/L,相对值为 0.22%。在这种情况下,组内标准差自由度为 100。重复性标准差可以由 M_{within} 计算[公式(C.5)]:

$$s_r = \sqrt{M_{\text{within}}} = \sqrt{1.63} = 1.28 \text{ IU/L} \quad \text{.....(C.5)}$$

这反映出重复性标准差相对较大,不足以证明不均匀性不低于 0.22%。根据参考文献[64]的推荐,按照公式(C.6),RM 生产者因此采用更加保守的方法来计算与不均匀性相关的不确定度,表示为 u'_{bb} 。

$$u_{\text{bb}} = \sqrt{\frac{M_{\text{within}}}{n_0}} \cdot \sqrt{\frac{2}{\nu_{M_{\text{within}}}}} = \sqrt{\frac{1.63}{6}} \sqrt{\frac{2}{100}} = 0.196 \text{ IU/L} \quad \text{.....(C.6)}$$

该表达式可表示为单元内标准差估计的标准不确定度。参考文献[16]建议使用由以上公式计算的 s_{bb} 和 u'_{bb} 中较大者。因此,将 0.196 IU/L 作为与单元间变异相关的不确定度。

C.3 使用随机化区组设计进行均匀性评估

C.3.1 背景和数据

河水中阴离子有证标准样品项目研制中,进行了单元间均匀性研究。这个实例使用硫酸根离子的

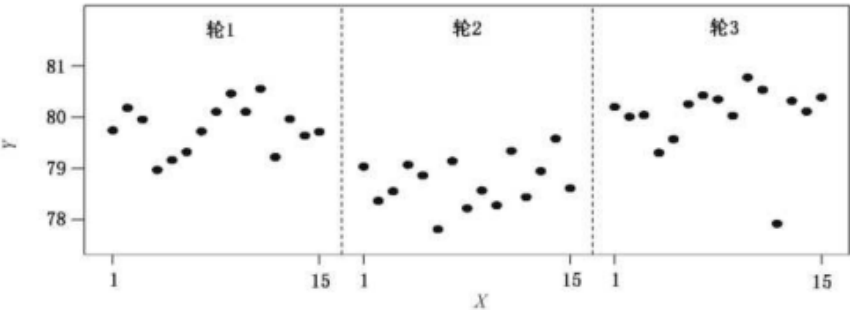
数据。研究采用随机区组设计,将 15 个单元的 RM 分别在三种不同的测量轮中测量一次。每轮中各个单元测量顺序都是随机的,确保各个单元在每轮以不同的次序出现。按照 RM 的单元编号排列出各单元硫酸根离子浓度数据,如表 C.4。

表 C.4 硫酸根的均匀性数据

单位为毫克每升

单元	轮 1	轮 2	轮 3
15	79.737	79.036	80.196
127	80.177	78.367	80.005
199	79.950	78.555	80.040
248	78.970	79.070	79.299
309	79.164	78.865	79.567
375	79.315	77.810	80.250
439	79.720	79.142	80.424
482	80.102	78.217	80.342
571	80.457	78.569	80.022
662	80.105	78.280	80.770
689	80.550	79.333	80.530
775*	79.215	78.442	77.919
841	79.96	78.947	80.313
911	79.637	79.574	80.107
973	79.710	78.614	80.385

注：随机的轮顺序在此没有显示。
* 单元 775 为一个离群值,被舍弃,见正文。



标引符号说明：
X —— 单元序号；
Y —— 硫酸根离子质量浓度 (mg/L)。
注：垂直虚线标识出不同的测量轮(表中轮 1 到轮 3)。本图中单元数 1-15 与表 C.4 中 15 个单元数对应。

图 C.1 候选标准样品中硫酸根的观测结果

表 C.4(图 C.1)中所有数据清晰显示出不同的测量轮对结果的影响,其中包括测量轮 3 中的离群值(结果显示是单元序号为 775 的样品),该离群值因错误的色谱峰所导致。剔除离群值方法有剔除单点值或者剔除受这个值影响的 RM 单元。在这种情况下,为保证总体平衡和简化数据分析,该单元的所有三个重复值都被剔除,留下 14 个单元是技术有效的数据。

C.3.2 评定

对表 C.4 中的数据进行了无重复的双因子方差分析(ANOVA),方差分析结果如表 C.5 所示,证实了测量轮对结果有非常显著的影响。

表 C.5 硫酸根数据的双因子方差分析

变异	自由度 (ν)	平方和 (SS)	均方 (MS)	F 值 (F value) ^a	Pr(>F) ^b
单元	13	3.065 1	0.235 8	1.347 6	0.249 6
轮	2	15.409 3	7.704 6	44.036 4	4.48×10^{-9}
残差	26	4.549 0	0.175 0		
^a F 统计量(均方除以残差均方)。					
^b p 值(零效应方差零假设下 F 统计量观测值的超出概率)。					

单元内和单元间方差 s_w^2 和 s_{bb}^2 按照 7.7.4 所示采用公式(C.7)和(C.8)进行计算:

$$s_w^2 = s_r^2 = M_w = 0.175\ 0 \quad \dots\dots\dots (C.7)$$

式中, M_w 为单元内均方, 给予 $s_w = 0.418\ 3\ \text{mg/L}$ 。

$$s_{bb}^2 = \max\left(\frac{M_A - M_w}{n_0}, 0\right) = \max\left(\frac{0.235\ 8 - 0.175\ 0}{3}, 0\right) = 0.020\ 3 \quad \dots\dots\dots (C.8)$$

式中, M_A 为单元间均方, n_0 为每个单元观测的次数(与这个简单区组设计的轮数一致)。这使得 $s_{bb} = 0.142\ 4\ \text{mg/L}$ 。

注 1: 通过对比, 对相同数据组进行单因子方差分析得到了相同的单元间均方(0.235 8), 但同时残差均方值却升至 0.712 8。这将错误地导致单元间方差 s_{bb}^2 估计值为零值, 低估了与均匀性有关的不确定度, 或者如果基于单元内各项进行均匀性补偿, 将会导致严重高估。

注 2: 这里描述的是随机区组设计, 其中每个 RM 单元出现在每个测量轮中, 是一个全因子设计或者是交叉设计的例子。交叉设计的双因子方差分析不同于 B.2 中描述的嵌套设计的双因子方差分析。

C.4 均匀性评估: 对线性趋势进行校正的简单设计

C.4.1 背景和数据

一个威士忌同系物有证标准样品的研制项目中, 进行了单元间均匀性的研究。采用单一气相色谱法对 35 个 RM 单元进行了 2 次重复的测试, 以随机顺序重复观测。表 C.10 列出了均匀性研究的数据和线性拟合后的残差(见 C.4.3)。表格中, 单元号对应充装顺序, 位置号指测量轮顺序中的位置。

按照 7.7.1, 数据可以以单元顺序和轮顺序绘图。轮顺序图见 C.2, 可见缓慢的上升趋势, 这源于仪器漂移, 图中直线是最小二乘法进行线性拟合得到的。线性回归的结果如表 C.6 所示。P 值(表中的“Pr(>|t|)”)足够小, 显示(斜率)对回归有显著影响。

表 C.6 甲醇对轮顺序号的回归统计

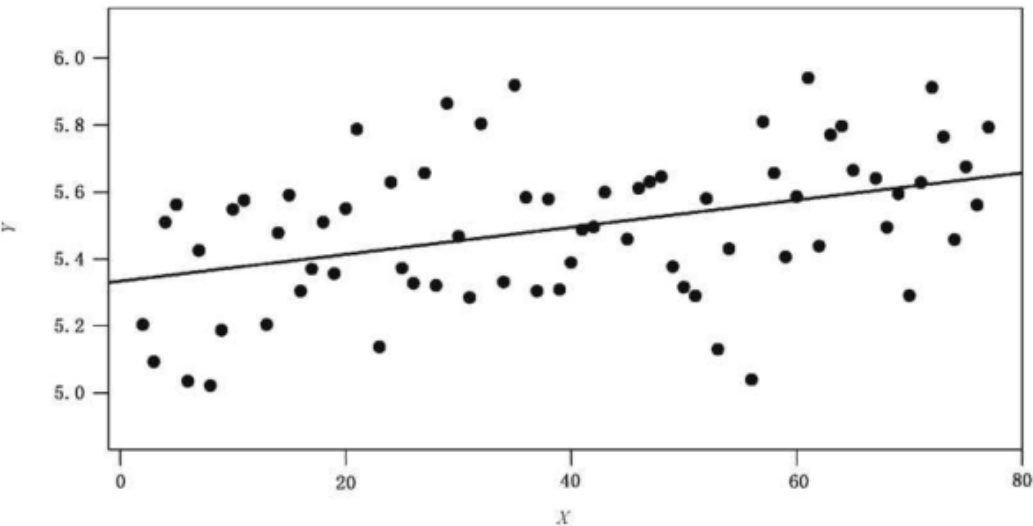
	估计值 ^a	标准误差 ^b	<i>t</i> 值 ^b	Pr(> <i>t</i>) ^c
截距	5.333 7	0.049 8	107.15	<2×10 ⁻¹⁶
斜率	0.004 0	0.001 1	3.69	0.000 45

^a 系数和标准误差保留小数点后四位。

^b *t* 统计量(估计值除以标准误差)。

^c *p* 值(斜率为 0 的零假设下观测到的统计量绝对值的超出概率)。

该趋势还没有强到需要拒绝实验轮次影响而重新分析,对图的直观检查发现可以通过线性拟合得到合理的近似值。因此,可以通过统计处理消除趋势的不利影响。有两种方法可行:一种是用经典方差分析进行修正,另一种是(仅用于结果比对)用同时包含了趋势和单元效应的混合效应模型的结果。



标引符号说明：
X —— 轮序列号；
Y —— 甲醇浓度 (mg/100 mL)。
注：实线是数据线性拟合后的结果，“轮序列号”与表 C.10 中的“位置”相对应。

图 C.2 按轮顺序绘制候选威士忌 RM 中甲醇的结果图

C.4.2 方法一：用经典方差分析对趋势进行校正

第一步从数据中减去线性趋势,计算后得到残差数据如表 C.10 所示。

注：残差相当于与线性模型中截距和位置(斜率)项的相减结果,由于随后的方差分析假设了单元效应和随机误差的可加性,因此减去位置(斜率)项并为截距保留未修正的数据是同样有效。

然后,把单元作为组因子,对残差进行单因子方差分析。趋势校正数据的原始方差分析表如表 C.7 所示。

表 C.7 威士忌样品残差均匀性的原始方差分析

变异	自由度 (ν)	平方和 (SS)	均方 (MS)	F 值 (F value) ^a	Pr(>F) ^b
单元间	34	1.764 6	0.051 90	1.714 6	0.058 65
单元内	35	1.059 5	0.030 27		
^a F 统计量(单元间均方除以单元内均方)。					
^b p 值(单元间无差异的零假设下观测值超出 F 统计量的概率)。					

然而,由于表 C.10 中的残差来自包含斜率和截距项的线性模型,表 C.7 中的方差分析表有点不正确,因为方差分析假设仅减去总平均值,因此作以下轻微修正:

从方差分析表中单元内自由度中减去额外的自由度,考虑斜率项的自由度。

通过将平方和除以修正后的自由度来重新计算残差均方。修正后的方差分析表(包括相应计算)如表 C.8 所示。

虽然对单元内均方的校正微小,但单元间项使用均方之间的差值,因此校正有时比要求的差值要大。

如果使用更复杂的模型进行趋势校正,则对单元内自由度的校正宜考虑所用模型中参数的数量。通常,如果模型中有 q 个参数,则从组内自由度中减去 $q-1$ 。

表 C.8 校正线性模型自由度的方差分析表

变异	自由度 (ν)	平方和 (SS)	均方 (MS)	F 值 (F value) ^a	Pr(>F) ^b
单元间	34	1.764 6	0.051 90	1.714 6	0.058 65
单元内	(35-1)=34	1.059 5	1.059 5/34=0.031 16		
^a F 统计量(单元间均方除以单元内均方)。					
^b p 值(单元间无差异的零假设下观测值超出 F 统计量的概率)。					

单元间和单元内标准差 s_{bb} 和 s_w 的测定按 7.7.3 公式(C.9)计算:

$$s_{bb}^2 = \max\left(\frac{MS_b - MS_w}{n_0}, 0\right) = \max\left(\frac{0.051\ 90 - 0.031\ 16}{2}, 0\right) = 0.010\ 4\ \text{mg}/100\ \text{mL} \cdots (\text{C.9})$$

得出单元间标准差 s_{bb} 为 0.101 8 mg/100 mL,单元内的标准差 s_w 是单元内均方的开方根,结果是 0.176 5 mg/100 mL。

比对发现,上文数据分析中得到的没有进行线性趋势校正的 s_{bb} 和 s_w 的估计值分别是 0.092 8 mg/100 mL 和 0.201 6 mg/100 mL,稍低估了单元间标准差,而高估了单元内标准差。

C.4.3 方法二:结合轮次趋势进行分析

在统计学术语中,仅受到单一轮次趋势影响的模型称为“固定效应”模型,而单元间效应被认为是随机的。要同时考虑并估计固定效应和随机效应时使用“混合效应”模型(见参考文献[26])。“混合效应”模型不仅估计随机效应的方差,特别是单元间效应,还估计固定效应的系数,比如趋势的斜率和截距。该方法可以适当调整以补偿自由度的损失。由于关注的单元内和单元间方差尽可能小,可以使用限制极大似然估计法(REML)(见 7.7.6)来估计。这就需要合适的统计软件。在下面的例子中,开源统计软

件 R 已载入 nlme 包^[65]。由于模型的技术参数和计算过程完全依赖于所使用的软件,因此只需关注结果。表 C.9 给出了各部分的参数以及一些简要介绍。

- 注意,软件给出的结果与 C.4.2 的经典计算结果比较接近。主要优点包括以下方面。
- 合适的软件计算时只需要较少的操作。在这种情况下,表 C.9 所显示的每项结果需要单独的命令,并且没有后续的计算。与 C.4.2 经典方法中的线性拟合、方差分析表计算、修正后的方差分析表,以及人工计算的标准差等步骤的操作相反。
 - 因为模型同时包含了单元间的随机影响和线性趋势的影响,而线性模型能更好地处理斜率和轮次的混杂效应,则可能会导致没有包含单元间效应的线性回归结果会偶尔包含过高或过低的估计结果。
 - REML 法能尽可能减少方差估计的偏差,并且保证是非负的;用经典方法可能给出负的方差估计结果^[67],而这就需要进一步处理数据。

表 C.9 混合效应模型软件的输出结果

软件输出结果	说明
模型	数据输入和拟合方法的简要说明
适用于 REML 方法的线性混合效应模型 数据:meoh 对数-限制-似然:5.537 651	数据集命名为“meoh”; 对数限制似然法结果是最优拟合模型返回的拟合值
固定效应:甲醇~斜率(轮次)	“甲醇~斜率”表示模型的固定效应部分,表示甲醇对轮次的线性拟合,与经典线性回归是一样的。模型系数可以与表 C.6 中的系数比较,尽管它们并不总相同
截距 斜率 5.328 8 0.004 173	
随机效应:公式~1 单元	随机效应以每个单元简单的随机影响建模
单元间标准差:0.104 2 单元内标准差:0.175 5	以标准差表示随机效应
观测值数:70 组数:35	与随机效应相关的组的大小的信息

在表中单元间标准差提供了 s_{bb} 的估计值,单元内标准差相当于 s_w 。

表 C.10 候选威士忌 RM 中甲醇浓度的结果

位置	单元	甲醇 mg/100 mL	残差 mg/100 mL	位置	单元	甲醇 mg/100 mL	残差 mg/100 mL
2	1 421	5.203 2	-0.138 6	9	224	5.186 6	-0.183 6
3	896	5.093 2	-0.252 7	10	687	5.548 9	0.174 7
4	1 482	5.510 2	0.160 3	11	337	5.575 5	0.197 2
5	406	5.562 5	0.208 5	13	602	5.203 9	-0.182 5
6	875	5.035 2	-0.322 8	14	105	5.478 1	0.087 7
7	1 263	5.426	0.063 9	15	1 482	5.591 3	0.196 8
8	787	5.022 3	-0.343 8	16	1 200	5.303 9	-0.094 6

表 C.10 候选威士忌 RM 中甲醇浓度的结果 (续)

位置	单元	甲醇 mg/100 mL	残差 mg/100 mL	位置	单元	甲醇 mg/100 mL	残差 mg/100 mL
17	406	5.370 7	-0.031 9	48	956	5.646 3	0.118 3
18	658	5.510 3	0.103 7	49	813	5.377 6	-0.154 5
19	1 368	5.355 8	-0.054 8	50	602	5.315 6	-0.220 5
20	1 368	5.550 8	0.136 1	51	1 337	5.288 8	-0.251 4
21	296	5.787 5	0.368 8	52	506	5.581	0.036 8
23	985	5.137 1	-0.289 7	53	1 129	5.129 8	-0.418 5
24	687	5.628 6	0.197 7	54	1 337	5.430 2	-0.122 1
25	1 163	5.371 9	-0.063	56	1 421	5.039 3	-0.521 1
26	506	5.327 1	-0.111 9	57	337	5.810 1	0.245 6
27	9	5.657	0.214	58	190	5.657	0.088 5
28	1 507	5.320 8	-0.126 3	59	1 309	5.407 2	-0.165 4
29	9	5.864 6	0.413 5	60	224	5.586 9	0.010 3
30	1 129	5.467 7	0.012 5	61	729	5.940 4	0.359 7
31	985	5.284 7	-0.174 5	62	296	5.439 4	-0.145 3
32	956	5.804 1	0.340 8	63	1 200	5.770 3	0.181 5
34	658	5.331 3	-0.140 1	64	875	5.796 1	0.203 3
35	383	5.919 3	0.443 9	65	53	5.665	0.068 1
36	1 263	5.584 1	0.104 6	67	105	5.640 6	0.035 6
37	813	5.303 9	-0.179 6	68	151	5.494 9	-0.114 1
38	151	5.578 9	0.091 3	69	1 074	5.594 8	-0.018 3
39	1 309	5.308 4	-0.183 2	70	1 507	5.290 3	-0.326 8
40	383	5.389 4	-0.106 3	71	1 074	5.628 2	0.007
41	1 036	5.487 3	-0.012 4	72	564	5.912 5	0.287 3
42	441	5.496 4	-0.007 4	73	1 036	5.765 1	0.135 9
43	441	5.599 2	0.091 4	74	1 163	5.458 2	-0.175 1
45	896	5.458 7	-0.057 2	75	787	5.675 7	0.038 4
46	729	5.611 9	0.092	76	564	5.561 4	-0.08
47	53	5.630 1	0.106 1	77	190	5.793 6	0.148 2
“位置”相当于测量轮顺序；QC 观测值已被删除。 “单元”相当于 RM 充装序列号。 “甲醇”相当于从 4 个机构获取的有关特性值。 “残差”通过线性拟合数据得到残差；残差在这张表中被截断至小数点后 4 位。							

附录 D

(资料性)

测量不确定度评定

描述测量不确定度评定的主要文件是 ISO/IEC 测量不确定度表述指南(“GUM”)^[4], ISO/IEC Guide 98-3:2008 第 8 章中提供了合成不同来源的不确定度分量的详细程序。该程序可以如下概括。

- a) 用数学表达式表示的特性值与特性值依赖的所有输入量之间的关系,称之为测量模型,模型宜包括可能对不确定度有显著贡献的所有的量。
- b) 根据一系列数据的统计分析或其他方法确定所有输入量的值。
- c) 通过对数据的统计分析或其他方法(如参考证书或专业判断)评定所有输入量的标准不确定度。
- d) 评定任何输入量之间的协方差。
- e) 计算特性值(X),即待认定的特性值。
- f) 从标准不确定度和与输入量相关的协方差来确定特性值的合成标准不确定度。
- g) 在规定的置信水平(通常为 95%)确定包含区间。在对称分布的情况下,包含区间可以用扩展不确定度的形式来表示,通过以特性值相关的标准不确定度乘以包含因子 k 来获得。
- h) 特性值宜与其包含区间一并报告,通常以区间的上限和下限表示;为了不确定度的传播,还宜给出与特性值相关联的标准不确定度;在特性分布是对称的情况下,该包含区间可以表示为扩展不确定度 U 和包含因子 k 。

在许多情况下,可以遵循如上所述的方法。然而,有些情况下,如 GUM^[4]中讨论的那样,宜选择其他方法,这样的情况包括:

- 描述特性值与输入量之间关系的模型没有确切的数学表达式;
- 通过合成和传播不确定度公式得到的线性近似显然是无效的。

对于非线性测量模型,宜确认使用不确定度传播律的有效性。如有疑问,宜采用 GUM 补充 1^[5]中所述的蒙特卡洛方法。

当认定一组特性有共同影响量时(例如,标准值均来自同一贮存溶液稀释而来的一组不同浓度的系列校准标准溶液),不同值之间的相关性可能会非常大,有必要向最终用户提供协方差信息。在这种情况下,可以使用多元不确定度评定方法来获得与一系列特性值相关的协方差矩阵。GUM 补充 2^[6]提供了这种方法的细节。

如何评定单一不确定度分量的详细内容,包含在 GUM^[4]及其补充文件中。在许多情况下,为利用现有的数据,可以定义组合的不确定度分量,如确认研究的结果,如参考文献[51]第 7 章所述。

参 考 文 献

- [1] ISO Guide 33, Reference materials—Good practice in using reference materials, International Organisation for Standardisation (ISO), Geneva, 2015, pp.31
- [2] ISO Guide 31, Reference materials—Contents of certificates, labels and accompanying documentation, International Organisation for Standardisation (ISO), Geneva, 2015, pp.10
- [3] ISO/IEC Guide 99, International vocabulary of metrology—Basic and general concepts and associated terms (VIM), International Organisation for Standardisation (ISO), Geneva, 2007, pp.92
- [4] ISO/IEC Guide 98-3, Uncertainty of measurement—Part 3: Guide to the expression of uncertainty in measurement (GUM: 1995), International Organisation for Standardisation (ISO), Geneva, 2007, pp.120
- [5] ISO/IEC Guide 98-3/Suppl.1, Uncertainty of measurement—Part 3: Guide to the expression of uncertainty in measurement (GUM: 1995)/Supplement 1: Propagation of distributions using a Monte Carlo method, International Organisation for Standardisation (ISO), Geneva, 2007, pp.88
- [6] ISO/IEC Guide 98-3/Suppl.2, Uncertainty of measurement—Part 3: Guide to the expression of uncertainty in measurement (GUM: 1995)/Supplement 2: Extension to any number of output quantities, International Organisation for Standardisation (ISO), Geneva, 2007, pp.73
- [7] Doc. BCR/48/93, Guidelines for the production and certification of BCR reference materials, European Commission, Brussels, 1994
- [8] ISO 2859 (all parts)²⁾, Sampling procedures for inspection by attributes, International Organisation for Standardisation (ISO), Geneva, 1999
- [9] ISO/REMCO Position Paper, The need for assessment of commutability of reference materials, International Organization for Standardization (ISO), Geneva, 2014, http://isotc.iso.org/livelink/livelink/fetch/-8854933/8854951/8854960/279217/Commutability_document_final.pdf?no-deid=16787892&vernum=-2
- [10] ISO 13528, Statistical methods for use in proficiency testing by interlaboratory comparison, International Organisation for Standardisation (ISO), Geneva, 2015, pp.89
- [11] Box G.E.P., Hunter J.S., Hunter W.G. Statistics for Experimenters; Design, Innovation, and Discovery, Second edition, J. Wiley and Sons, 2005
- [12] Ellison S.L.R., Grimes N., Fardus F., Ellison, M.P.L., Use of a replicated Latin square design in a homogeneity test for high purity organic melting point standards, 2012, Accreditation and Quality Assurance, Vol.17, p.283 to 290
- [13] Van der Veen A.M.H., & Linsinger T.P.J., Lamberty A., & Pauwels J Uncertainty calculations in the certification of reference materials 4.Characterisation and certification, 2001, Accreditation and Quality Assurance, Vol.6, p.290 to 294
- [14] ISO 5725-3, Accuracy (trueness and precision) of measurement methods and results—Part 3: Intermediate measures of the trueness of a standard measurement method, International Organisation for Standardisation (ISO), Geneva, 1994, pp.25
- [15] Searle S.R., Casella G., McCulloch C.E. Variance components, J. Wiley and Sons, 2006
- [16] Linsinger T.P.J., Pauwels J., Van der Veen A.M.H., Schimmel H., Lamberty A. Homoge-

2) 该系列标准已部分修订。

neity and Stability of Reference Materials, 2001, Accreditation and Quality Assurance, Vol.6, p.20 to 25

[17] Ingamells C. O., & Switzer P. A proposed sampling constant for use in geochemical analysis, 1973, Talanta, Vol.20, pp.547 to 568

[18] Van der Veen A.M.H., Linsinger T.P.J., Lamberty A., Pauwels J. Uncertainty calculations in the certification of reference materials. 2. Stability study, 2001, Accreditation and Quality Assurance, Vol.6, p.257 to 263

[19] Lamberty A., Schimmel H., Pauwels J. The study of the stability of reference materials by isochronous measurements 1997, Fresenius Journal of Analytical Chemistry, Vol.360, p.359 to 361

[20] ICH Q1D, Bracketing and matrixing designs for stability testing of new drug substances and products, International Conference on Harmonisation, 2002

[21] Tydeman M.S., & Kirkwood T.B.L. Design and analysis of accelerated degradation tests for the stability of biological standards II. A flexible computer program for data analysis, 1984, Journal of Biological Standardization, Vol.12, pp.207 to 214

[22] Bremser W., Becker R., Kipphardt H., Lehnik-Habrink P., Panne U., Töpfer A. Stability testing in an integrated scheme, 2006, Accreditation and Quality Assurance, Vol.11, p.489 to 495

[23] Becker R., Buge H.-G., Bremser, W., Nehls I. Mineral oil content in sediments and soils: comparability, traceability and a certified reference material for quality assurance, 2006, Analytical and BioAnalytical Chemistry, Vol.385, p.645 to 651

[24] Koch M., Bremser W., Köppen R., Töpfer A., Nehls I. Development of two Certified Reference Materials for Acrylamide Determination in Foods, 2009, Journal of Agriculture and Food Chemistry, Vol.57, p.8202 to 8207

[25] West B.T., Welch K.B., Galecki A.T. Linear Mixed Models: A Practical Guide Using Statistical Software. Chapman and Hall/CRC, 2006

[26] Pinheiro J., & Bates D. Mixed-Effects Models in S and S-PLUS. Springer, 2000

[27] Linsinger T.P.J., & Emons H.E. Characterization of reference materials: proposal for a simplification of the options listed in ISO Guide 34, 2013, Accreditation and Quality Assurance, Vol.18, p.149 to 152

[28] EURACHEM/CITAC Guide. Traceability in Chemical Measurement. EURACHEM, 2003

[29] ISO 6865, Animal feeding stuffs—Determination of crude fibre content—Method with intermediate filtration, International Organisation for Standardisation (ISO), Geneva, 2000, pp.10

[30] ISO 6142-1, Gas analysis—Preparation of calibration gas mixtures—Part 1: Gravimetric method for Class I mixtures, International Organization for Standardization (ISO), Geneva, 2015, pp.39

[31] Milton M.J. T., Vargha G. M., Brown A. S. Gravimetric methods for the preparation of standard gas mixtures, 2011, Metrologia, Vol.48, pp.R1 to R10

[32] Van der Veen A.M.H., & Hafner K. Atomic weights in gas analysis, 2014, Metrologia, Vol.51, pp.80 to 86

[33] Ruhkin A. L., & Vangel M. G. Estimation of a Common Mean and Weighted Means Statistics, 1998, Journal of the American Statistical Association, Vol.93, pp.303 to 308

[34] CCQM Guidance note, Estimation of a consensus KCRV and associated Degrees of Equivalence, Consultative Committee for Amount of Substance (CCQM), Bureau International des Poids et Mesures (BIPM), Sèvres, Paris, 2013

[35] Cox M. The evaluation of key comparison data, Metrologia, 2002, Vol.39, pp.589 to 595

- [36] Van der Veen A.M.H. Determination of the certified value of a reference material appreciating the uncertainty statements obtained in the collaborative study, in Ciarlini P., Cox M.G., Filipe E., Pavese F., Richter D. *Advanced Mathematical and Computational Tools in Metrology V*, Singapore, World Scientific, 2001, p.326 to 340
- [37] Pauwels J., Lamberty A., Schimmel H. The determination of the uncertainty of reference materials certified by laboratory intercomparison, 1998, *Accreditation and Quality Assurance*, Vol.3, pp.180 to 184
- [38] Paule R.C., & Mandel J. Consensus Values and Weighting Factors, 1982, *J Res Natl Inst Stand Technol*, Vol.87, p.377 to 385
- [39] Saito T., Ihara T., Koike M., Kinugasa S., Fujimine Y., Nose K., Hirai T. A new traceability scheme for the development of international system-traceable persistent organic pollutant reference materials by quantitative nuclear magnetic resonance, 2009, *Accreditation and Quality Assurance*, Vol.14, p.79 to 86
- [40] Saito T. Efficient production of reference materials of hazardous organics using smart calibration by nuclear magnetic resonance, 2011, *Accreditation and Quality Assurance*, Vol.16, pp.421 to 428
- [41] Westwood S. Mass balance method for the SI value assignment of the purity of organic compounds, 2013, *Analytical Chemistry*, Vol.85, pp.3118 to 3126
- [42] ICH Q3a, Harmonised tripartite guideline Q3A(R2) Impurities in new drug substances, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) ICH, [Online] 2006, [Cited: 11 March 2013] <http://www.ich.org>.
- [43] EUR 22504, Certification of a reference material of purified genomic DNA from *Campylobacter jejuni* (NCTC 11351) IRMM-448, 2007
- [44] European Pharmacopoeia, Council of Europe, Strasbourg, Eighth Edition, 2014
- [45] Haynes W.M. *CRC Handbook of Chemistry and Physics*, CRC Press, Boca Raton, Ninety fifth Edition, 2014 [ed.]
- [46] McNaught A.D., & Wilkinson A. [ed.], *IUPAC Compendium of Chemical Terminology*, International Union of Pure and Applied Chemistry, 2014. XML on-line corrected version created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins, ISBN 0-9678550-9-8
- [47] 2002/657/EC, 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]
- [48] ISO 11843-1, Capability of detection—Part 1: Terms and Definitions, International Organization for Standardization (ISO), Geneva, 1997, pp.10
- [49] Josephs R., Koeber R., Bernreuther A., Schimmel H., Ulberth F. EUR Report 21202 EN; The certification of the mass fraction of aflatoxin M1 in whole milk powders Certified Reference Material ERM®-BD282 (zero level) ERM®-BD283 (low level) ERM®-BD284 (high level), Joint Research Centre, European Commission, 2005
- [50] Hunter R. S. Accuracy, precision and stability of new photoelectric colour-difference meter, *J. Opt. Soc. Am.* 1948, 38 p.1094A
- [51] Ellison S. L. R., & Williams A. eds. *Eurachem/CITAC guide: Quantifying Uncertainty in Analytical measurement*, 2012, Available from www.eurachem.org.
- [52] ISO/IEC 17025, General requirements for the competence of testing and calibration laboratories, International Organisation for Standardisation (ISO), Geneva, 2005, pp.28

- [53] Levenson M.S., Banks D.L., Eberhardt K.R., Gill L.M., Guthrie W.F., Liu H.K., Vangel M.G., Yen J.H., Zhang N.F. An approach to combining results from multiple methods motivated by the ISO GUM, 2000, *J. Res. Natl. Inst. Stand. Technol.*, Vol. 105, pp. 571 to 579
- [54] Toman B., & Possolo A. Laboratory effects models for interlaboratory comparisons, 2009, *Accreditation and Quality Assurance*, Vol. 14, pp. 553 to 563
- [55] Van der Veen A.M.H., & Pauwels J. Uncertainty calculations in the certification of reference materials. 1. Principles of analysis of variance, 2000, *Accreditation and Quality Assurance*, Vol. 5, p. 464 to 469
- [56] Sokal R.R., & Rohlf F.J. *Biometry*. Freeman, New York, Third Edition, 1995
- [57] John P.W.M. *Statistical design and analysis of experiments*. Macmillan, New York, 1971
- [58] Snedecor G.W., & Cochran W.G. *Statistical Methods*. Iowa State University Press, Eighth Edition, 1989
- [59] Searle S.R. *Linear Models*. John Wiley, New York, 1971
- [60] Griepink B., Maier E., Muntau H., Wells D.E. The certification of the contents of six chlorobiphenyls (Nos 28, 52, 101, 108, 153 and 180) in dried sludge (CRM 392), European Commission Joint Research Centre, Luxembourg, 1990. Certification report; EUR 12823
- [61] Van der Veen A.M.H., Linsinger T.P.J., Pauwels J. Uncertainty calculations in the certification of reference materials 2. Homogeneity study, 2001, *Accreditation and Quality Assurance*, Vol. 6, p. 26 to 30
- [62] Montgomery D.C. *Design and Analysis of Experiments*. John Wiley & Sons, Hoboken, NJ, 2012
- [63] Schiele F., Siest G., Moss D.W., Colinet E. The certification of the catalytic concentration of gamma-glutamyltransferase in a reconstituted lyophilized material (CRM 319), European Commission, Joint Research Centre, 1986. Certification report. EUR 10628
- [64] Linsinger T.P.J., & Pauwels J. Lamberty, A., Schimmel, H., Van der Veen, A.M.H., Siekmann, L., Estimating the uncertainty of stability for matrix CRMs, 2001, *Fresenius Journal of Analytical Chemistry*, Vol. 370, p. 183 to 188
- [65] Pinheiro J., & Bates D., DebRoy S., Sarkar D and the R Development Core Team, *Linear and Nonlinear Mixed Effects Models*. R package version 3.1-103, 2012
- [66] R Development Core Team. *R: A language and environment for statistical computing*, Vienna, Austria, R Foundation for Statistical Computing, 2012
- [67] Rice J.A. *Mathematical statistics and data analysis*, Second edition, Duxbury Press, Belmont, 1995
- [68] ISO Guide 30, *Reference materials—Selected terms and definitions*, International Organisation for Standardisation (ISO), Geneva, 2015, pp. 8
- [69] IUPAC. *Compendium of Chemical Terminology*, 2nd ed. (the Gold Book), Blackwell Scientific Publications, Oxford, 1997
- [70] Milton M.J. T. The mole, amount of substance and primary methods, *Metrologia*, 2013, Vol. 58, p. 158 to 163
- [71] ISO 15189, *Medical laboratories—Requirements for quality and competence*, International Organisation for Standardisation (ISO), Geneva, 2012, pp. 53
- [72] Bettencourt Da Silva R., & Williams A. *Eurachem/CITAC Guide: Setting and Using Target Uncertainty in Chemical Measurement*, (1st ed. 2015), Available from www.eurachem.org
- [73] ISO 7870-2, *Control charts—Part 2: Shewhart control charts*, International Organisation

for Standardisation (ISO), Geneva, 2013, pp.43

[74] Ellison S.L.R., Burke S., Walker R.F., Heydorn K., Månsson M., Pauwels J. Uncertainty for reference materials certified by interlaboratory study; Recommendations of an international study group, 2001, Accreditation and Quality Assurance, Vol.6, p.274 to 277

[75] ISO 5725-2, 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, International Organisation for Standardisation (ISO), Geneva, 1994, pp.42

[76] ISO 16269-4, Statistical interpretation of data—Part 4; Detection and treatment of outliers, International Organisation for Standardisation (ISO), Geneva, 2010, pp.54

[77] Lee P.M., Bayesian Statistics; An Introduction, John Wiley & Sons, Chichester, UK, 2004

[78] Ellison S.L.R. Homogeneity studies and ISO Guide 35; 2006, 2015, Accreditation and Quality Assurance, Vol.20, p.519 to 528

[79] De Bièvre P., Dybkær R., Fajgelj A., Hibbert D.B. Metrological traceability of measurement results in chemistry; Concepts and implementation (IUPAC Technical Report), 2011, Pure and Applied Chemistry Vol.83 (10), pp.1873 to 1935

[80] ISO/TR 16476, Reference materials—Establishing and expressing metrological traceability of quantity values assigned to reference materials, International Organisation for Standardisation (ISO), Geneva, 2016, pp.23

[81] LGC/VAM, Meeting the Traceability Requirements of ISO 17025, An Analyst's Guide, LGC, 2003

[82] ISO/TR 79, Reference materials—Examples of reference materials for qualitative properties, International Organisation for Standardisation (ISO), Geneva, 2015, pp.38

[83] Magnusson B., & Örnemark U. eds. Eurachem Guide; The Fitness for Purpose of Analytical Methods—A Laboratory Guide to Method Validation and Related Topics, (2nd ed. 2014), ISBN 978-91-87461-59-0, Available from www.eurachem.org.

[84] ISO 17034, General requirements for the competence of reference material producers, International Organisation for Standardisation (ISO), Geneva, 2016, p.24
