

JJF

中华人民共和国国家计量技术规范

JJF××××—202×

标准物质研制（生产）中的定性分 析通用技术要求

General Technical Requirements for Qualitative Analysis in
Reference Material Production

（征求意见稿）

202X—XX—XX 发布

202X—XX—XX 实施

国家市场监督管理总局发布

标准物质研制（生产）中的定性分析通用技术要求

JJF xxxx-202x

General Technical Requirements for Qualitative Analysis in Reference Material Production

归口单位：全国标准物质计量技术委员会

主要起草单位：

参加起草单位：

本规范委托全国标准物质计量技术委员会负责解释

本规范主要起草人：

参加起草人：

国家标准物质计量技术委员会征求意见稿

目 录

引言.....	III
1. 范围.....	1
2. 规范性引用文件.....	1
3. 术语.....	1
4. 定性特性.....	1
5. 定性分析.....	2
5.1 概述.....	2
5.2 基于来源.....	2
5.3 基于测量.....	3
5.4 多种方法联用.....	4
6. 溯源性的建立.....	5
6.1 概述.....	5
6.2 计量溯源性.....	5
6.3 追溯性.....	6
7. 不确定度的评估.....	6
7.1 概述.....	6
7.2 测量不确定度.....	7
7.3 定性特性值的可信度.....	8
8. 统计学方法的选择.....	9
8.1 概述.....	9
8.2 选择因素.....	9
9. 定性特性标准物质的均匀性评估.....	9
9.1 概述.....	10
9.2 实验设计.....	10
9.3 统计方法.....	11
10. 定性特性标准物质的稳定性评估.....	14
10.1 概述.....	14
10.2 实验设计.....	15

10.2 统计方法.....	16
11. 定性特性标准物质的互换性评估.....	18
11.1 概述.....	18
11.2 评估方法.....	19
11.3 互换性声明.....	20
附录 A DNA 和蛋白质定性标准物质	21
附录 B 定性特性值可信度的表达	24
附录 C 常见统计学程序介绍	32
附录 D 定性特性标准物质的均匀性和稳定性研究示例.....	36

国家标准物质计量技术委员会征求意见稿

引言

根据 ISO 33406《定性特性标准物质的生产方式》，定性特性包含了《国际通用计量学基本术语》中明确的标称特性和序值特性。尽管相关术语和概念尚处在不断的认识和发展中，且与现行 ISO/IEC 指南 99《国际计量学词汇——基本和通用概念及相关术语》（VIM）和 JJF1001《通用计量术语及定义》存在差异，但国家标准物质的研制（生产）中，逐渐出现特性涉及标称特性和序值特性的标准物质，对定性特性开展准确定值或对定量特性有关被测量等开展可靠定性鉴别的需求日益增加。

本规范参考 ISO 33406 制定，且相关要求不低于 ISO 33406。与 ISO 33406 的主要区别在于：

- 1) 根据我国计量领域的标准物质研制（生产）范围和特性值类型，对定性特性标准物质的定值途径进行限定，定值方法中应至少包括一种基于测量的可靠方法，且定值结果具有可靠的计量溯源性保障；
- 2) 增加规范的适用性，针对定量特性标准物质日益复杂的被测量或被测特性等的定性鉴别或检查要求，提供相关通用指导；
- 3) 根据以上考虑和国家计量技术规范起草要求，对文件名称、结构和内容等作相应调整。

本规范是对 JJF1342、JJF1343 和 JJF1854 的补充，如无特别指明，以下内容适用于本规范：

- 1) JJF1342 对标准物质研制（生产）的通用要求、结构要求、资源要求和管理要求；
- 2) JJF1343 对标准物质定值、赋值、均匀性和稳定性评估的通用原则和要求；
- 3) JJF1854 对测量结果计量溯源性的通用原则和要求。

本规范为首次发布。

标准物质研制（生产）中的定性分析通用技术要求

1. 范围

本规范给出了标准物质研制（生产）中开展定性分析的基本原则、方法和要求，适用于基于测量的标准物质定性特性，尤其是标称特性的定值和赋值，以及开展定量特性有关标准物质研制（生产）时，对原料来源、归属、物种身份、真实性、品质等被测量或被测特性信息（定性特性）等的鉴别或检查。

2. 规范性引用文件

本规范引用下列文件：

JJF 1001 通用计量术语及定义

JJF 1005 标准物质通用术语和定义

JJF 1342 标准物质研制（生产）机构通用要求

JJF 1343 标准物质的定值及均匀性、稳定性评估

JJF 1186 标准物质证书和标签要求

JJF 1218 标准物质研制报告编写规则

JJF 1854 标准物质计量溯源性的建立、评估与表达计量技术规范

ISO 33406 定性特性标准物质的生产方式（Approaches for the production of reference materials with qualitative properties）

凡是注日期的引用文件，仅注日期的版本适用于本规范；凡是不注日期的引用文件，其最新版本（包括所有的修订单）适用于本规范。

3. 术语和定义

JJF1001、JJF1005 和 JJF1343 中界定的有关术语和定义适用于本规范。

4. 定性特性

定性特性（qualitative property）又称分类特性（categorical property），包括标称特性和某些序值特性。定性特性可具有两个或多个可能得值。根据定性特性值的不同，材料可被分为两个或更多个类别，同一类别中的所有材料具有相同的

定性特性值。对于标称特性，特性值之间唯一可以进行的比较是相同或不同。对于序值特性，特性值之间还可按相对等级顺序进行比较，即一个特性值等于、低于或高于另一个特性值。本规范重点关注标称特性。

5. 定性分析

5.1 概述

5.1.1 定性分析结果可以定性特性值的形式表达，也可以定量特性值的形式表达。例如：对于颗粒形状或依据亨特（Hunter）系统得到的颜色的定性分析，可基于对量的测量实现，而被测量则可由操作或程序定义。对于颜色，还可考虑采用吸收或反射光谱进行测量。

5.1.2 许多定性特性无法进行有意义的量化。例如：物质归属、生物物种或动物性别。

5.1.3 对于量的测量，被测量相同是通过计量溯源性建立测量结果计量可比性的前提，因此，需要对被测量进行必要的定性确证。如：纯度标准物质及为标准物质特性值提供量值溯源的纯物质原料的纯度值应建立在对纯物质可靠确证的基础上；当原料的品质，如某些杂质的存在，影响标准物质的可靠定值和赋值时，也应开展必要的定性鉴别。

5.1.4 为满足预期用途，定性分析应仅在针对该预期用途，具有足够高的把握时进行。

5.1.5 定性分析包括以下途径：

- 1) 基于一个或多个定性测定；
- 2) 基于来源，见 5.2；
- 3) 基于测量，见 5.3；
- 4) 基于多种方法的组合，见 5.4。

5.1.6 本规范中标准物质定性特性的定值，应至少包括一种或一种以上基于测量的方法，且定值结果具有可靠的计量溯源性保障。

5.2 基于来源

5.2.1 定性分析可基于对材料来源，即出处的了解进行。此时，应获取文件或其

他证据，证明从材料来源到最终成品包装的不间断证据链，所有证据应在材料的整个生命周期内保存。

5.2.2 应在材料的处理（如采样、均质化、包装、储存）过程中制定措施，以确保材料来源的有效性，并防止被其他材料污染。

注：对于生物材料，出处可包括亲本证据或来自真实样本（包括标准物质）的连续培养证据。

5.2.3 宜有额外证据来确认材料的真实性，参见 5.3 和 5.4。

5.3 基于测量

5.3.1 定性分析可基于一种或多种方法的定量测量结果来实现，如核磁共振（NMR）波谱的化学位移和峰面积，或颜色、熔点、相对分子质量等的组合。

5.3.2 定性分析的结果可通过将材料的一组测量结果与这些测量结果的预定义接受标准进行比较得到，如熔点范围、与参考 DNA 序列的相似性等。

例：有机高分子材料可以通过与红外（IR）参考谱图的比较来识别，采用的预定义接受标准可包括：与参考谱图中所有峰的频率偏差在 3 cm^{-1} 以内；与参考谱图的相对峰强度偏差在 5% 相对吸光度范围；参考谱图中的所有峰在该有机高分子材料的谱图中均存在；该有机高分子材料谱图中的所有峰在参考谱图中均存在。

注：材料的定性分析有时也可通过专家判断来（如石棉纤维、组织病理学检查）进行。该判断通常基于观察和比较得出，包括与判定标准的比较和与相关用户群体认可的标准物质、凭证标本或参考数据的比较，例如：可以用于基于与凭证标本的比较来识别植物物种。基于观察的专家判断属于 5.1.5 中的“基于一个或多个定性测定”。

注 2：序列的相似性可以报告为“一致性百分比”，即两个序列之间实现特征匹配的百分比。

5.3.4 材料的轻微不均匀性和不稳定性不一定改变定性分析的结论。均匀性和稳定性评估的目标是材料的适用性，即是否仍然能够进行明确的定值或鉴别。

5.3.5 对于所选择的方法，不同材料也可表现出相同的定量特性。因此，有关材料来源和加工处理步骤的信息至关重要。

5.3.6 在标准物质项目策划中，应根据材料的预期用途，明确规定对该材料定性鉴别信息或目标定值特性的需求，以及相关预定义接受标准或判定标准。

例：对于 DNA 标准物质，根据其预期用途，可能仅需要物种鉴别声明，也可能需要完整序列或甲基化程度的信息。

5.3.7 预定义接受标准的来源可包括国际公认标准汇编，如目标物质的参考数据集。此类文献信息可能会发生变化，因此，应在标准物质研制报告等文件中清晰说明用于定性鉴别或定值的具体标准，可以为一组值，也可以为具有日期的引用文件。

5.3.8 在确定预定义接受标准时，应比较各类文献数据，确定报告值的范围，为每个被测量建立反映所用信息范围和可靠性的标准，并将其文件化。应优先选择经过同行评审的参考数据。

5.3.9 结果与预定义接受标准的比较，同样也适用于基于测量表征的其他定性特性。

5.4 多种方法联用

5.4.1 该方式特别适合于小到中等分子质量的特定化学物质，宜选择能够对其不同特性进行检测的多个方法开展联合确证。常用方法包括熔点、相对分子质量、紫外（UV）、红外（IR）、核磁共振（NMR）和质谱等。

5.4.2 结合原料和加工步骤的信息，以及取样和运送到标准物质研制生产机构的信息，方法的组合应足以实现对定性特性的可靠分析。

注 1：所需方法的性质和数量取决于存在潜在类似材料或物质的可能性，以及材料来源和处理步骤等有关信息的充分性。

注 2：也可根据材料的预期用途，运用适当的感官分析方法。

例 1：当使用来自公认保藏中心的细菌标准菌株时，其附带文件中可能需要包括宏观特征（典型培养基上的特征）、微观特征（例如革兰氏染色的特征、球菌或杆菌鉴定）、表型特征（生化反应中的特征）和其他特征（例如血浆凝固酶测定、抗氧化酶测试、血清学测试）的信息。

例 2：从研究机构获得的植物种子如果仅具有形态特征，可能不足以支持该植物的分类；可能需要进一步的证据或证明文件。

5.4.3 如果存在详细的已发布标准，则所需方法可限定为标准中规定的方法。

5.4.4 所有方法和程序应经过适当确认，且结果满足溯源性要求。如果可能，应在对材料进行定性测定的同时，测定适当的控制样品。

5.4.5 每项测量和检查的结果应与目标物质的接受标准进行比较。比较宜采用已发布的程序（如果有）。如果没有此类规定程序，在考虑合成测量不确定度和规定值的情况下，测量结果不宜与任何规定值之间存在差异。如果结果与发布标准

一致，则材料鉴别的不确定度可以忽略。

注 1：判断积累的测量或材料来源的信息是否足以支持定性特性的鉴别或定值，在某种程度上是主观的。同行评审有助于提高对赋值结果的信心。

注 2：相似性量度有助于使用定性和定量方法的组合来开展定性特性的鉴定或定值。ISO/TR 79 给出了相似性量度的示例。

5.4.6 在适用的情况下，应对定性特性标准物质的纯度进行适当的评估或说明，以支持定性特性的鉴定或定值。

6. 溯源性的建立

6.1 概述

6.1.1 本规范中的溯源性包括：

- 1) 计量溯源性：适用于所有量的测量；
- 2) 追溯性：适用于基于来源的定性分析，以及基于与人工制品（如作为实物标准的有证标准物质）、权威数据（如已发表并受到广泛认可的样照及其相关说明）等参照对象的比较来开展的定性分析。

6.1.2 建立计量溯源性和追溯性的目的是确保定性分析结果的可比性、一致性。

6.1.3 对于标准物质定性特性的赋值，应按照 JJF1186 要求，在标准物质证书中给出相关溯源性声明。标准物质研制报告应能按照 JJF1218 要求，提供溯源性的详细信息，并保留有关溯源性证明。

6.2 计量溯源性

6.2.1 对于基于测量（如熔点或 pH）的定性分析和规定了量化测试条件（如时间、温度和浓度）的定性分析，相关测量或测试应符合 GB/T 27025 中的计量溯源性要求。

例 1：在使用光谱方法进行结构定性时，对光谱信号位置有关的波长或频率标尺进行校准或检查。

例 2：在需要使用规定浓度试剂的定性分析测试中，用于制备试剂的设备经适当的校准，以确保浓度的可靠性。

6.2.2 当标准物质定性特性的赋值涉及测量或量化测试条件时，应参考 JJF1186 和 JJF1854，在标准物质证书中包含相关计量溯源性声明。

6.3 追溯性

6.3.1 当标准物质定性特性的赋值须依赖并追溯至规定的标准物质（包括可用于比较的身份可靠的样品或经独立定值的样品）、参考数据等可靠的参照对象时，应在标准物质证书中就所采用的参照对象作出说明。

6.3.2 开展定量特性有关标准物质研制（生产）时，对原料来源、归属、真实性、品质等被测量或被测特性信息等的定性鉴别，应在标准物质研制报告中说明所采用的参照对象。

例：对于草药的鉴定，可通过与国家保藏库或档案馆中保存的草药标准样品进行显微比较来进行。

6.3.3 对于基于来源的定性分析：

1) 应保留有关材料可追溯性的全部文件，包括材料的来源、所有权变更、避免污染或材料被替换的程序。

2) 应在标准物质证书或研制报告中就相关鉴别依据（包括材料的来源）作出说明。

6.3.4 当需要通过验证性的测量或测试手段来支持基于来源的定性鉴别时，应参照 6.2 和或 6.3.1。

7. 不确定度的评估

7.1 概述

7.1.1 对被测特性的每一次赋值均存在不确定度。JJF 1059.1 中定义的测量不确定度仅适用于量值。

7.1.2 对于定性特性值的不确定度，本规范采纳以下原则：

1) 定性特性值的不确定度可用可信程度，即“可信度”表示。可信度可通过定量方式表达，如似然比或定性特性可能值集上的概率分布来表示，也可用定性方式，即序值标度来表示，如“最可信”、“非常可信”、“中等可信”。

2) 针对标准物质的预期用途，其定性特性值的结果应具备足够的可信度。标准物质研制（生产）机构应清晰说明建立可信度的依据，记录和提供可信度评估的支持性信息。

3) 标准物质的定性特性值应附对其可信度的声明,并在标准物质证书中提供。可信度应尽可能以定量的方式给出声明。当以定性方式表述时,应避免造成对定性特性值可信程度的误解。

4) 开展定量特性,如化合物纯度标准物质研制(生产)时,对原料来源、归属、真实性、品质等被测量或被测特性信息等的定性鉴别,通常提供序值标度的可信度即可。相关信息应在标准物质研制报告中提供。

注:附录 B 给出了关于定性特性值可信度表述的更多信息。

7.2 测量不确定度

7.2.1 JJF 1059.1 中规定的测量不确定度以两种方式影响定性特性的鉴别或赋值:

1) 对测试条件进行控制的不确定度,如时间、温度或长度,当涉及量值测量或需对测试条件进行量化控制时,该不确定度至关重要。

2) 临界测量相关的测量不确定度,可能会影响对误判率(假阳性率或假阴性率)的估计,例如:当依赖超过某个阈值的测量结果进行分类时。

7.2.2 当定性分析依赖于测量或量化的测试条件时,应确保测量不确定度足够小,以不对定性特性的鉴别结果或标准值的可信度产生显著影响。具体要求包括:

1) 建立并以文件形式规定影响测试结果的条件和容差范围,并进行有效控制和记录。

2) 证明不确定度足够小,不会对测试结果和定性特性值产生显著影响。

注:由于低的误判率难以确定,相对可行的方法是证明条件的重大变化,如超出容差限以上三倍不确定度或中间测量结果三倍不确定度时,对误判率的影响有限,或影响未检出。

7.2.3 当定性分析依赖于测量值与一个或多个限值的比较时,可使用测量不确定度,结合对结果分布的适当假设来估计误判的概率。

7.2.4 当标准物质的定性特性认定值依赖于一个或多个量值的测量,认定值以定量特性值的形式表达(参见 5.1.1)时,原则上应参照 JJF1059.1,评估定量特性值的测量不确定度。

7.2.5 在定性特性标准物质证书中,不需要报告与临界测量结果有关的测量不确定度信息。

7.3 定性特性值的可信度

7.3.1 标准物质研制（生产）机构应提供足够的信息，以允许用户评估标准物质定性特性值的可信度是否满足其需求。

7.3.2 对标准物质定性特性值的可信度进行量化表述，有利于帮助用户判断该标准物质是否满足其预期用途。

注：量化的可信度声明可以是置信概率或置信水平。

7.3.3 帮助用户对定性特性值的可信度开展评估的信息可包括：

- 1) 赋值的依据，例如：来源、测试、专家结论；
- 2) 材料来源的说明；
- 3) 为了将材料受到污染或因偶然因素发生替换的可能性降至最小，所采取的有关程序的描述；
- 4) 赋值相关定性测试方法的原理和性能的信息；
- 5) 用于支持赋值的参考数据的信息；
- 6) 可信度的定性或文字表述；
- 7) 可信度的定量表述。

注 1：定性分析的性能指标参见 B.2。

注 2：可信度的定量表示参见 B.4。

7.3.4 如果提供测试方法性能有关的信息，其呈现方式应避免引起对定性特性值可信度的误解。

例：使用核酸测序方法开展鉴别时，理论上偶然匹配的概率低于 10^{-50} ，但用户据此推断所赋值的错误概率仅为 10^{-50} 是有风险的，实际上的概率可能高得多。因此，针对缺少经验的用户，可能需要在测试性能数据的相关性或解读方面作出警示性建议，例如使用序值标尺来给出证据强度（参见 B.1）。

7.3.5 可信度的定性表达可以是对可信度等级的简要解释性说明或表明赋值依据的简短陈述。如果给出了可信度的等级，则应向用户提供完整的等级标度。应基于清晰的准则开展定性分析，这些准则应明确且特别适用于特定的场景，并按照 7.1.2 中的要求，在标准物质证书或研制报告中说明。

例：对于银杏叶，与生物物种识别相关的不确定度以在某等级标度上的文字表述来表示（参见 B.1），可声明如下：赋值结果的可信度分为弱、中等、强和非常强，测试结果为所赋值提供了非常强的支持。

注：有时，可信度的水平来源于数值，例如：似然比或概率（参见 B.4）。

7.3.6 可信度的定量表达可以是目标特性在可能值集上的概率分布，也可以是对该分布在可能值集上分散程度的概括表述。也可通过分类错误（假阳性、假阴性）率、灵敏度和特异性来表达（参见表 B.2），或通过受试者工作特征曲线（ROC 曲线）表达。

例：某些 DNA 测序技术通过整数质量分数（integer quality scores）来表征对占据特定位点碱基鉴别的可信度，四种可能的碱基（A、C、G 或 T）各对应一个质量分数。

注：对所赋定性特性值可信度的量化评估参见 B.4。

8. 统计学方法的选择

8.1 概述

定量特性赋值以及均匀性、稳定性和互换性评估相关的统计学方法较为完善，但广泛应用于定性分析的统计学方法极为有限。因此，本规范第 8、9、10 章对特定目的下的统计学方法给出通用指导，也可采用其他充分适用的统计学方法。

8.2 选择因素

定性分析统计学方法的选择应考虑以下因素：

1) 定性测试返回的信息通常少于返回数值结果的定量测量。因此，应优先选择给出与待测定性特性相关定量信息的测量方法，而不是仅给出定性信息的方法。例如对于生物物种特性的稳定性评估，可以通过指示生物物种存在的特定蛋白质的浓度来进行，而不是测试蛋白质存在与否的定性测试程序。

2) 当定性测试方法具有显著的误判率时，将难以区分误判是由测试失败引起还是由标准物质缺陷或变化引起。因此，应尽可能选择可靠性高的测试方法进行定值、均匀性评估和稳定性评估。

3) 标准物质预期是充分均匀和稳定的，不符合预期的不一致性结果极为罕见，甚至不存在，例如：当使用可靠的测试方法开展某特性的稳定性检查时，所有测试样品及重复测量结果均为阳性，无阴性结果。因此，应选择适合处理不一致性响应比例低或为零情况的统计学方法。

9. 定性特性标准物质的均匀性评估

9.1 概述

9.1.1 依据 JJF1342 要求, 为确保标准物质的适用性, 应对其在最终包装形式中的均匀性开展评估, 包括每批次标准物质的单元间均匀性和单元内均匀性。对于单元内均匀性, 通常需要考虑最小取样量。

9.1.2 应评估标准物质全部定性特性值的均匀性。

9.1.3 定性特性标准物质的均匀性评估应采用直接针对所赋定性特性值的测试方法和标准物质证书中与预期用途有关的任何其他方法。

注: 每个特性可能具有不同程度的均匀性分布。对所赋定性特性值开展的均匀性评估, 并不一定能保证标准物质用户关注的所有特性的均匀性, 即使相关测试方法属于同一类别。

9.1.4 当通过量化测量手段来支持定性特性标准物质的赋值时, 宜选择相关定量特性的测量方法, 依据 JJF1343 开展均匀性评估。

9.1.5 如果每个标准物质单元的定性特性为独立赋值, 则无需评估该特性的单元间均匀性。

9.2 实验设计

9.2.1 JJF1343 提供了定性特性标准物质均匀性研究的通用实验设计, 包括获取代表性标准物质单元的抽样策略。相关设计和抽样策略直接适用于采用定量特性的均匀性实验研究, 也可应用于定性特性的均匀性评估。

注: 更多指导参见以下文献: Eurachem/Citac Guide Assessment of the performance and uncertainty in qualitative chemical analysis. Edited by R. Bettencourt da Silva and S.L.R. Ellison 2021. ISBN 978-0-948926-39-6。

9.2.2 当采用给出定性结果的测试方法开展均匀性研究时, 还需考虑以下因素:

1) 宜增加用于研究的标准物质单元数量, 以便为不合格单元率的估计提供足够窄的置信区间 (见 9.2.3), 或在采用计数检验抽样方案时, 宜根据所选择的抽样方案确定抽样单元数量 (见 9.3.3)。

2) 如果所选测试方法的误判率可以基本忽略不计 (即所有合格标准物质单元在任何单次测试中都会得到合格的结果, 或不合格单元会得到不合格的结果), 则可以省去重复测试。即可对每个抽取的标准物质单元仅进行一次测试。

3) 用于置信概率分析的最小样本数量取决于提供定性测试结果的判定标准的类型。

注：对于情况 2)，除非有额外的信息（如测试方法的变动性），单元间不均匀性与单元内不均匀性或测试方法的偶然失误是无法区分的。但是如若单元间无差异，则可解读为缺少不均匀性的证据。

9.2.3 对于标准物质的定性特性，用于均匀性研究的单元数量宜根据 ISO 2859 系列标准或类似指南中介绍的计数抽样检验原则确定。

注：对于定性特性，均匀性实验研究可能仅限于检查非预期的粗大不均匀性（例如大于10%的缺陷），同时提供其他用于支持均匀性声明的信息，如有关材料来源、加工等的信息。对于计数抽样检验，在不合格单元占比不高的情况下，如果仅依赖于抽样检验，则抽样数量会非常高，例如，要想基于一个或多个观测到的缺陷，以高置信度检测到1%的不合格率，需要检查大约300个随机抽取的单元（此数量基于批量远大于300单元的假定，根据1%不合格率和95%置信概率下的统计学功效计算得到）。对典型的标准物质批量，通常并不现实。

9.3 统计方法

9.3.1 原则和程序

9.3.1.1 可采用以下两种通用方式：

1) 确认不合格标准物质单元的比例不超过选定的临界值。该方式对每个标准物质单元重复测试或单次测试均适用。

2) 检验不合格响应概率的单元间显著差异。该方式在每个标准物质单元执行重复测试时适用。

9.3.1.2 9.3.1.1 中的方式 1) 可通过计算不合格标准物质单元比例的上置信限来实现，见 9.3.2；也可采用计数检验原则，见 9.3.3。

9.3.1.3 9.3.1.1 中的方式 2) 可通过多种统计程序实现。9.3.4 和 9.3.5 给出了两种适合的程序。

9.3.2 标准物质单元不合格率的置信区间

9.3.2.1 应设定标准物质单元不合格率的判定标准，若不符合该标准要求，则标准物质样品作废或需进一步加工。

9.3.2.2 对于 9.2 中的任何实验设计，都应计算标准物质不合格单元率的单侧置信上限，以形成均匀性评估的结论。应采用针对低不合格率的置信区间计算程

序。合适的程序包括“精确”皮尔森-克洛帕（Pearson-Clopper）二项式区间和威尔逊（Wilson）评分区间。如果每个标准物质单元进行了重复测试，则应基于被判定为不合格的标准物质单元数量，而不是响应值的总数来计算置信区间。在未观测到不合格的特殊情况下，标准物质单元不合格率的简单近似区间为 $[0, 3/n]$ ，其中 n 为测试标准物质单元的数量。

9.3.2.3 计算出的置信限应与 9.3.2.1 中的判定标准比较，如果置信限小于或等于所选判定标准，则该材料具有足够的均匀性，可被接受。

注 1：该程序检验的是对判定标准的符合性，以保证标准物质单元不合格率低于一定的水平，达到该水平则需考虑再次加工。

注 2：附录 C.2 介绍了 Wilson 区间的计算方法。

注 3：附录 D.1.4 给出了均匀性研究中 Wilson 区间的计算示例。

9.3.3 计数检验

9.3.3.1 GB/T 2828.1 和 GB/T 2828.4 给出了计数检验的抽样和验收方案。这些方案用于重复批的检验，对标准物质批量重复生产相对频繁的情况尤为有用。基本原则如下：

1) 设置一个“接收质量限”（AQL），即当连续一系列批次提交验收时，可接受的标准物质单元不合格率的最差过程平均值。

注：AQL 作为抽样方案的参数，不应和用于反映标准物质生产过程操作水平的过程平均相混淆，该过程平均应优于 AQL，以避免过度拒收。

2) 选择“检验水平”。检验水平表明检验程序的严格程度。GB/T 2828.1 和 GB/T 2828.4 建议大多数情况采用“正常”水平或声称质量水平（DQL）为 0。

注：对于重复批检验，ISO 2859-1 根据最近通过或未通过检验的批次数量来调整检验水平。

3) 根据生产批的标准物质单元数量（ISO 2859-1 中的“批量”）选择抽样方案。对于单个生产批，适用单次抽样方案。

4) 抽取所需数量的代表性标准物质单元样本，并测试样本中的所有标准物质单元。

5) 如果没有发现缺陷（不合格标准物质单元），则该批予以接收。如果检测到缺陷，则将缺陷数量与抽样方案中的接收和拒收数量进行比较。如果缺陷数量大于或等于拒收数量，则该批予以拒收。

注 1：对于该程序予以拒收的批次，不合格标准物质单元比例显著大于约 95%置信水平下的 AQL。因此，实际的缺陷 RM 单元比例可能远高于 AQL。

注 2：不合格单元比例的上 95%置信限可以使用附录 C.2 中的程序计算。

注 3：对于批量小于 250 的标准物质，GB/T2828.11 和以下统计学文献可提供更多指导：Montgomery D.C. Introduction to Statistical Quality Control, John Wiley & Sons, Hoboken, NJ, 7th edition, 2013. ISBN 978-1-118-14681-1。

例：对于大约 2,000 个标准物质单元的生产批，没有关于均匀性的先验信息。AQL 设定为 1.0（即 1.0%不合格品）。检验水平设定为正常（GB/T 2828.1 中的水平 II）。参考 GB/T 2828.1 中的单次验收抽样方案，批量大小落入“样本量字码”K（批量 1,201 至 3,200 个）。对于正常检验的单次抽样方案，要求检查 125 个标准物质单元。AQL 为 1.0 的可接收数量为 3（见 GB/T 2828.1），拒收数量为 4。因此，如果不合格单元数量小于 3 或更少，该批予以接收；如果有 4 个或更多，则拒收。

9.3.3.2 在使用单次抽样方案的重复批生产和检验中，如果连续五个批次中的两批未通过检验，则检验水平应提高到“加严”，如果连续五个批次通过加严检验，则可恢复到正常检验水平。GB/T 2828.1 还提供了基于据累积分数的放宽检验规则。

注：调整检验水平的详细规则见 GB/T 2828.1。

9.3.3.3 如果误判率已知，或者有目标特性值的其他信息可用，则 GB/T 2828.4 中的方法可显著减少样品数量。

9.3.4 二维列联表分析

9.3.4.1 对于多个标准物质单元的重复测试，双向列联表按标准物质单元的阳性（不合格）和阴性（合格）结果进行计数。统计学检验方法，如卡方检验或 Fisher 精确检验（阴性响应数较少时使用），可被用于检验不同单元间阴性响应概率的显著性差异。

9.3.4.2 对于均匀性测试，如果检验统计量超过 95%临界值，或相应的 p 值小于 0.05，则认为标准物质是不均匀的。

注 1：附录 C.1 介绍了适用于二维列联表的卡方检验程序。

注 2：附录 D.1.3 介绍了使用卡方检验对均匀性研究进行分析的示例。

9.3.5 分类方差分析（CATANOVA）

分类方差分析（CATANOVA）或针对序值响应的序值方差分析（ORDANOVA），可用于检验不同单元间不一致结果概率的显著性差异。这

两种程序适用于多个标准物质单元重复测试的实验设计。

注 1：通用的详细计算程序见文献：Light, R.J. and MARGOLIN B.H. An analysis of variance for categorical data. Journal of the American Statistical Association. 1971, 66(335), 534–544. doi: 10.1080/01621459.1971.10482297。应用于均匀性测试的详细计算程序见文献：Gadrich, T. and BASHKANSKY, E. ORDANOVA: Analysis of ordinal variation. Journal of Statistical Planning and Inference. 2012, 142(12), 3174–3188. doi: 10.1016/j.jspi.2012.06.004。

注 2：关于双因素 CATANOVA 和双因素 ORDANOVA 的进一步信息，见文献：

1) Marmor, Y.N. and BASHKANSKY, E. Processing new types of quality data. Quality and Reliability Engineering International. 2020, 36(8), 2621–2638. doi: 10.1002/qre.2642;

2) Gadrich, T., KUSELMAN, I., ANDRIĆ, I., Macroscopic examination of welds: Interlaboratory comparison of nominal data. SN Appl. Sci. 2020, 2, 2168.

10. 定性特性标准物质的稳定性评估

10.1 概述

10.1.1 依据 JJF1342 要求，应评估标准物质定性特性值的稳定性，包括：确保运输条件适当的运输稳定性、确保储存条件适当的长期稳定性，以及终端用户使用期间的稳定性。

10.1.2 标准物质定性特性值的稳定性，应根据测试方法针对该特性值的响应稳定性来评估。

10.1.3 当需要开展稳定性实验评估时，应采用以下一种或多种方式：

1) 确定定性测试方法对所研究定性特性有预期正确响应的单元比例的变化。

2) 研究可指示材料是否发生降解的定量特性。

注 1：稳定性研究的方式和特性是对用户有用的信息。

注 2：10.2 给出了采用定性测试方法或定量方法进行稳定性研究的细节设计。

10.1.4 当通过量化测量手段来支持定性特性标准物质的赋值时，宜根据 JJF1343 评估标准物质的稳定性，并选择相关定量特性的测量方法。

10.1.5 应通过标准物质证书向用户提供标准物质定性特性赋值结果在有效期内的稳定性的说明，以及有关标准物质的处理和使用方法。

注：标准物质的每个特性可能具有不同的稳定性。支持标准物质定性特性赋值的被测

特性的稳定性，并不一定能保证标准物质用户关注的所有特性的稳定性，即使测试方法属于相同类别。

10.2 实验设计

10.2.1 基本设计

10.2.1.1 JJF1343 提供了稳定性研究的通用实验设计，包括获取代表性标准物质单元的抽样策略。这些设计适用于采用定量特性开展的稳定性实验研究，也可应用于定性特性的稳定性评估。

10.2.1.2 最简单的定量或定性特性稳定性研究实验设计是将大量单元暴露于计划的储存（或其他）条件下，并以固定的时间间隔测试一定数量的单元，即 JJF1343 中的经典稳定性研究。与之密切相关的替代方案是同步设计，即将标准物质单元定期移动到参考条件下，所有测试推迟到研究结束时进行。对于定性特性，经典和同步设计都是在一组时间的每个时间点提供失效单元的比例估计。

10.2.1.3 可系统改变其他因素，如温度和湿度，以开展加速研究或测试额外因素的影响。

10.2.1.4 稳定性研究中选择的标准物质单元数量取决于实验设计、对于用户的潜在风险、研究时间内失效率的预期变化以及结果所需的可信度。作为参考，在没有先验稳定性证据的情况下，对于简单的经典和同步稳定性研究，宜每个时间点 10 个标准物质单元，或整个研究共 40 个标准物质单元。

注：在 95% 置信水平下，采用通过双侧、双样本比例检验，每组 10 个观测值是可靠识别失效概率从零变到 50% 的最小数量，可为在每个时间点以最少并可行的单元数量来检测出大的变化提供良好保障。

10.2.2 持续监测设计

当可以定期且非破坏性地对所有单元的状态进行监测时，可采用另一种稳定性研究实验设计，即将所有标准物质单元作为一个大组进行连续监测，并记录每个标准物质单元失效的时间，直至预定的最长时间。通过该方法可以得到一组已知的失效时间，以及一组已知失效时间大于研究时长的标准物质单元，继而可采用生存分析方法进行评估。10.2.5 给出了简要的介绍。

10.2 统计方法

10.2.3 列联表分析

10.2.3.1 稳定性实验设计中，每个时间点的失效和未失效标准物质单元数量可构成最简单的二维列联表，继而通过采用卡方或其他假设检验的列联表分析，对不同时间点失效比例的差异进行统计学检验。显著的差异则说明可能存在不稳定性。对于其他因素，如温度，可通过额外的交叉列表来处理。

10.2.3.2 应谨慎解释稳定性数据的列联表分析结果。列联表分析不考虑不同时间或温度点的排序或间隔信息，仅检验组间差异。因此，组间差异显著，并不能说明失效率随时间的持续升高。在检出显著性差异的情况下，宜进一步检查是否存在一致性趋势，例如绘制失效比例随时间变化图。

注 1：适用于二维列联表的卡方检验程序见附录 C.1。

注 2：列联表分析应用于稳定性研究的示例见附录 D.2。

10.2.4 失效概率随储存或暴露时间的回归分析

10.2.4.1 逻辑回归可以针对失效单元的计数或比例，也可针对单个失效。对于单个失效，通常以编码 0 代表未失效标准物质单元，1 代表失效或有缺陷的单元。对于最简单的逻辑回归模型，失效概率随时间的变化由公式（1）表述。

$$\ln\left(\frac{p}{p-1}\right)=b_0+b_1t \quad (1)$$

其中： p 为失效概率；

$\ln$ 为表示自然对数；

t 为特定保存条件下的持续时间；

b_0 和 b_1 为模型系数。

该模型类似于最简单的定量特性稳定性评估线性模型。

10.2.4.2 Probit 回归可作为逻辑回归的替代方法。Probit 回归采用公式（2）所示模型。

$$p=\phi(b_0+b_1t...) \quad (2)$$

其中： p 为失效概率；

ϕ 为累计正态分布函数；

t 为特定保存条件下的持续时间；

b_0 和 b_1 为模型系数。

10.2.4.3 实际上，逻辑回归和 Probit 回归在形式上相似。可基于以下因素选择：

- 1) 导致失效的潜在过程；
- 2) 模型对数据的拟合度；
- 3) 软件的可获得性。

注 1：模型对数据的拟合度可使用赤池（Akaike）信息准则（AIC）或贝叶斯信息准则（BIC）来评估。

注 2：逻辑回归可以推广应用到具有两个以上值的标称特性或序值特性，见参考文献第 8 章：Beauchamp C.R., Camara J.E., Carney J., Choquette S.J., Cole K.D., Derose P.C. et al. Metrological Tools for the Reference Materials and Reference Instruments of the NIST Materials Measurement Laboratory, NIST Special Publication 260-136 (2021 edition), National Institute of Standards and Technology, Gaithersburg, MD. doi: I

10.6028/NIST.SP.260-136-2021。

注 3：逻辑回归和 Probit 回归属于广义线性模型。有关模型的详细描述和及数据拟合计算方法见以下参考文献：

- 1) Agresti A. Categorical Data Analysis, John Wiley & Sons, Hoboken, NJ, 3rd edition, 2013. ISBN 978-0-470-46363-5;
- 2) Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. Applied Logistic Regression. John Wiley & Sons, Hoboken, NJ. 3rd Edition, 2013. ISBN 978-0-470-58247-3;
- 3) McCullagh, P. and Nelder, J. A. Generalized Linear Models, Chapman & Hall/CRC, London, UK, 2nd edition, 1989. ISBN 0-412-31760-5。

应用示例见参考文献：Pawitan, Y. In all likelihood: statistical modelling and inference using likelihood. Clarendon Press, Oxford University Press, 2001. ISBN 9780198507659。

注 4：当数据集中没有或只存在极少数失效（或几乎所有单元都失效）时，逻辑回归和 Probit 回归通常无法成功。在加速研究分析中，一些标准物质单元的暴露条件可以提高早期失效概率，因此，以上方法更为有用。

10.2.5 生存分析

7.2.5.1 生存分析最早应用于临床研究，是对累积失效进行的时间统计分析，有时也被称作可靠性分析，特别是在工程领域。生存分析的某些方法特别适用于 10.2.2 中持续监测设计有关的数据分析。鉴于生存分析涉及领域广泛，本规范不做详细介绍。表 1 对一些主要模型及其优缺点进行了简要总结，可分为参数方法（假设了失效概率随时间的特定模型）和非参数方法（不假设特定模型）。

7.2.5.2 所有生存分析模型都可以处理“右删失”数据，即仅已知失效时间大于研究持续时间的情况。所有模型都需要软件实现。

表 1 部分生存分析模型总结

模型	公式	优缺点
逻辑模型	见公式 (1)	简单模型，通常由统计软件提供，适用于单个观测或通过一组观测计算得到的比例。
指数模型 ^a	$S(t) = e^{-\lambda t}$	具有单一系数 λ 的简单“一阶”模型，提供随时间恒定的失效风险。
Weibull 模型 ^a	$S(t) = e^{-(\lambda t)^\alpha}$	比指数模型更灵活的模型，具有 λ 和 α 两个系数。对于 $\alpha > 1$ ，失效风险随时间增加，对于 $\alpha < 1$ ，失效风险随时间减少。
比例风险模型 ^b	$h_1(t) = \psi h_0(t)$	该模型用单一参数 ψ 来对两条生存曲线之间的关系进行建模，该参数量化了两种条件下失效风险的差异。该模型需要一个用于对比的对照条件，因此最适合研究条件变化的影响，例如不同的包装方式或更高的储存温度。该模型的主要优点是，当失效模型的形式未知时，可以比较失效风险。该模型通常应用于对整体进行持续监测而得到的数据。
^a 对于参数模型，模型形式以生存函数 $S(t)$ 给出，该生存函数描述了随时间 t 的生存概率。 ^b 对于比例风险模型，模型形式以风险函数 $h(t)$ 给出，假设标准物质单元保持未失效直至时间 t ，则该函数给出时间 t 的失效概率。		

11. 定性特性标准物质的互换性评估

11.1 概述

11.1.1 标准物质的互换性与其作为校准品或质量控制（QC）物质的应用能力有关。当标准物质的定值采用一种测量或测试程序（通常为参考程序）表征，常规样品的测试采用另一测量或测试程序，不同程序对不同类型测试材料的响应有显著差异时，这一点尤其重要。并非所有的标准物质，都需要进行评估互换性，但对于某些类别和用途的标准物质则非常重要，如：用于体外诊断的标准物质；用于定性测试程序中阳性或阴性控制的定性特性标准物质。

注 1：当已知测量程序对于标准物质基体和预期常规样品中的被测量均具有足够的特异性时，通常不需要评估互换性。

注 2：当标准物质，如基体标准物质，与常规样品的来源和处理方式相同时，通常不需要建立互换性。

11.1.2 如果材料的互换性对于特定的预期用途是必须的，而标准物质研制（生产）机构对标准物质适合该特定用途予以保证，则应进行互换性评估。

注 1：GB/T 19703（被 GB/T 21919 引用）要求对体外诊断程序的有证标准物质进行互换性评估。

注 2：更多信息见参考文献：ISO/TC334 Position Paper. The need for assessment of commutability of reference materials, International Organization for Standardization (ISO) [online]. Geneva, 2014[viewed 25 October 2023]. Available from [https:// committee .iso .org/ sites/ tc334/ home/references/ commutability .html](https://committee.iso.org/sites/tc334/home/references/commutability.html)。

11.2 评估方法

11.2.1 互换性评估宜包括以下一项或多项内容：

1) 考虑可能影响互换性的材料的所有物理、化学和生物学特性，包括特定的认定特性以及用于提高标准物质稳定性、均匀性或使用性能的任何加工或处理；

2) 查阅已发表的类似材料互换性的文献数据；

3) 拟定互换性评估实验计划。

11.2.2 定性特性标准物质互换性评估实验的目的是提供实验性证据，以证明对于预期用途条件下的检查程序，标准物质和典型的测试样品具有相同的相应。以下情况宜进行互换性实验研究：

1) 对材料特性的考虑提示缺少互换性可能会对标准物质的使用产生不利影响时；

2) 需要满足适用标准的要求时。

11.2.3 当被测特性只有一种可能的测试程序时，标准物质对该程序的响应足以符合预期。但当标准物质预期用于多种程序时，则宜确认其对常用程序的响应。当不同的程序基于不同的原理时，则宜确认标准物质对每种不同原理代表性程序的响应。

例：对于认定特性为特定过敏原成分的食品基体标准物质，可基于聚合酶链反应方法进行核酸序列测试，也可基于免疫或液相色谱串联质谱法（LC-MS/MS）进行蛋白质检测。该情况下，重要的是要确认对于常规食品检测中代表性的核酸和蛋白质检测方法，该标准物质具有正确的响应。

11.3 互换性声明

当需要提供互换性信息时，有关信息应充分，以供终端用户判断该标准物质是否适合特定用途而无需进一步确认，或在使用前还需要开展额外的确认。特别是，标准物质证书中应明确：

- 1) 是否进行了互换性研究；
- 2) 如已进行研究，给出该标准物质已被证明具有互换性的测试方法和不具互换性的测试方法；
- 3) 标准物质与常规测试样品之间可能会影响互换性的任何已知差异。

附录 A

DNA 和蛋白质定性标准物质（资料性）

A.1 DNA 标准物质

A.1.1 DNA 标准物质的定性特性值

A.1.1.1 DNA 标准物质一般可被赋予两种类型的定性特性值：一种是目标 DNA 链有关部分的核酸碱基序列——A（腺嘌呤）、C（胞嘧啶）、G（鸟嘌呤）和 T（胸腺嘧啶）；另一种是标准物质生产所源自的物种的身份。

A.1.1.2 DNA 测序技术的基础是对定量测量信号的获取和评估。评估后得到的标称特性结果以 DNA 序列的形式表示。有关测序数据质量评估的指南见 GB/T 43584.1。

A.1.1.3 DNA 标准物质标称特性的赋值取决于：

- 1) 对加工用原材料来源和纯度的控制；
- 2) 所采用的 DNA 测序技术或其他 DNA 技术手段（例如聚合酶链反应，PCR）。

A.1.2 来源和纯度控制

A.1.2.1 鼓励标准物质研制（生产）机构通过获取客观证据来确保用于标准物质加工的原材料是真实的，例如：

- 1) 从菌种资源库获得特定的微生物菌株，从中提取基因组 DNA。该微生物菌株附有确认材料真实性的分析证书；
- 2) 选择用于克隆的 DNA 片段，并严格控制克隆条件（例如采用同一不相容群的高拷贝载体进行单菌落培养）；
- 3) 采用已知身份和来源的动物或植物组织样本（例如由分类学家进行了形态学鉴定的物种、进行了遗传鉴定的已知 DNA 序列）。

A.1.2.2 标准物质的纯度可能发挥重要作用，应予以评估。宜评估潜在 DNA 污染对标准物质预期用途的影响，例如，来自宿主细菌细胞的残留基因组 DNA 或残留痕核糖核酸(RNA)分子可能会导致基于紫外吸光度的 DNA 定量偏移。同样，对于物种身份或真实性的确证，杂质也会影响测量结论。

A.1.3 其他核酸材料方面的应用

A.1.3.1 A.1.1 和 A.1.2 中的原则也可适用于其他核酸标准物质（例如 RNA）。

A.2 蛋白质标准物质

A.2.1 蛋白质标准物质的定性特性值

蛋白质分析涉及蛋白质组或单个蛋白质的鉴定、定量、纯度评估、表达、翻译后修饰的确定、诱导和转换（降解和再合成）。其中，蛋白质的鉴定是确定蛋白质的一级氨基酸序列，或根据与生物体已知基因标准型（模式标本）的基准 DNA 序列相关联的氨基酸残基，确定所关注的特定蛋白质的名称（或登录号）的过程。蛋白质鉴别作为蛋白质组学研究的重要环节，是针对所研究系统开展后续生物学解读的前提。

A.2.2 氨基酸序列的测定

A.2.2.1 对于单一蛋白质的标准物质，应采用可靠性已得到证明的技术来开展氨基酸序列的测定，例如：高分辨质谱（MS）结合艾德曼（Edman）降解。适当时，应将一级氨基酸序列的测定与基于基准 DNA 序列构建、经深度注释的蛋白质数据库进行比对。

注 1：深度注释的蛋白质数据库诸如：Entrez 蛋白质数据库、RefSeq 数据库和 UniProt 数据库（包含 Swiss-Prot 及作为补充的 TrEMBL）。

注 2：深度注释通常指数据库对蛋白质功能、结构域、翻译后修饰、变体和亚型的描述具有最小冗余水平。相关实例有牛（*Bos taurus*）、小鼠（*Mus musculus*）、酵母（*Saccharomyces cerevisiae*）、大肠杆菌（*Escherichia coli*）和做得最好的人类（*Homo sapiens*）。

A.2.2.2 可通过肽质量指纹图谱进一步验证蛋白质的一级氨基酸序列。在通过氨基酸分析来确定蛋白质一级序列时，无论采用何种技术或技术的组合，均应证明相对于已知序列的覆盖度，并注明该已知序列所对应的数据库及其版本，以上要求需体现在所有标准物质文件中。该覆盖度应与和样品同步分析的参考蛋白质的覆盖度相当。

A.2.3 复杂材料中单一或多种蛋白质的鉴定

A.2.3.1 对于复杂材料中的蛋白质鉴定，应采用可靠性已得到证明的技术，例如 LC-MS/MS，并通过检索序列数据库来进行谱图鉴定。所有必要的、与蛋白质鉴定有关的样本制备步骤（例如增溶介质、二硫键还原、游离硫残基处理、蛋白水

解酶)应形成文件。需要文件化的内容还包括:鉴定用的检索算法、软件版本号和/或发布日期(例如 Byonic、Comet、Mascot、MS Amanda、MS-GF+、Sequest)、所使用蛋白质数据库的来源及版本、鉴定用的数据库检索参数标准(例如质量范围、碎裂模式、质量分析器、分辨率)、允许的错切位点数量和非特异性切割位点数量、所考量的全部修饰类型(静态和可变)的清单,以及标准物质鉴定赋值过程中使用的任何其他参数。应记录用于判定谱图可作为蛋白质鉴定充分依据的接受标准(例如阈值得分或因子判据)。还宜记录肽和蛋白质水平上的错误发现率(FDR),包括任何用于评估鉴定(不)确定度或确定 FDR 的附加统计分析(例如诱饵库检索结果或其他统计和计算方法的结果)。

A.2.3.2 蛋白质鉴定赋值的可靠性受检索算法数量的影响,增加算法数量可以提高鉴定赋值的可信度。通过 LC-MS/MS 以及从头测序和/或结合多种方法的谱图鉴定,可对复杂材料中的蛋白质进行进一步确证。

A.2.3.3 由于蛋白质数据库和用于蛋白质鉴定的软件算法的更新,建议将用于标准物质中蛋白质鉴定赋值的数据(包括原始数据和处理过的数据,以及实验和技术有关元数据)存档在适当的、公开可用的数据存储库中(例如 PRIDE、MassIVE、Peptide Atlas、Panorama)。

附录 B

定性特性值可信度的表达（资料性）

B.1 总体原则

B.1.1 与定量特性值类似，定性特性值的赋值通常也具有不确定度。但是，对于定性特性，不确定度通常表达为可信度。可信度的表达可以是定量的，也可以是定性的。对于可信度的定性表达，通常采用反映可信度级别的序值标度。

B.1.2 可信度的定性评估与表达应基于明确且适用的标准并在相关证书和文件中体现。

—例：对于银杏叶的物种鉴定，其不确定度通过与序值标度的比较来评估和表达：

1) 最确信（有支持度极高且充分解析的系统发育树和/或将该物种与近缘物种区分的多个诊断核苷酸位点；有含包容性和排他性的多个样品（样品集）的数据；有多个独立基因区域的一致性数据）。

2) 非常确信（有支持度较好且充分解析的系统发育树和/或将该物种与近缘物种区分的几个诊断核苷酸位点；有含包容性和排他性的多个样品（样品集）的数据；有一个或多个独立基因区域的一致性数据）。

3) 确信（有支持度较好且充分解析的系统发育树和/或将该物种与近缘物种区分的一个或几个诊断核苷酸位点；有含包容性和排他性的几个样品（样品集）的数据；有一个或多个独立基因区域的大体一致的数据）。

4) 不明确（系统发育树解析不佳、支持度较差和/或缺少将该物种与近缘物种区分的诊断核苷酸；有含包容性和排他性的几个或多个样品（样品集）的数据；一个或多个独立基因区域的数据总体不一致）。

B.2 定性特性测试方法的性能指标

B.2.1 对于只返回一个标称特性结果的测试方法，其性能指标可借助实验评估来表征。评估中，测试若干已知标称特性值的样品，并记录测试结果。假设只有两种可能的测试结果（这里用阳性和阴性），可能出现的结果及计数由表 B.1 进行汇总。表 B.2 则描述了可由表 B.1 中的结果计数得到的常见性能指标。

注 1：表 B.2 中的性能指标最常用于二元测试方法，此类测试结果以“阳性”和“阴性”区分，例如分别对应单个特征的存在与缺失。但是，该类方法也可用于从大量互斥结果中判定某个特定值（例如特定的生物或化学物种）的存在的情况，此时，分类正确为真结果，分类

错误为假结果。这就使这些指标有可能应用于有两个以上值的情况。

注 2: B.3 款提供了有关标称特性一致性量度的信息, 可用来建立梯度化的一致性等级。

B.2.2 对于一个有效的测试方法, 错误响应计数 (n_{FP} , n_{FN}) 一般很低, 也可能偶然为零。为了避免错误响应率的严重低估, 对于观察到的计数, 可以使用置信上限或应用其他推论方法, 例如: “伪计数”这样的简单程序, 是对研究中每个观察到的计数加 1, 然后再计算比例。

注: 基于贝叶斯定理, 可以证明将 1 加到每个观察到的计数是合理的。假定从观察到的错误响应计数 n_F 和总体 $N = n_F + n_2 + \dots + n_n$ 个响应中估计得到的特定错误响应概率 p_F 的先验分布为均匀分布, 可以证明 p_F 的后验分布的均值为 $(n_F+1)/(N+n)$; 即在每个计数上加 1 后的比例。虽然对于较低的 n_p , 这种错误响应率的估计存在正向偏移, 但当计数为 0 的情况较为普遍时, 通常是更有用的方法。

B.2.3 表 B.2 中的任何性能指标都不能直接解释为某个特定标称值正确 (或不正确) 的概率。然而, 有几个指标可被解释为 (估计的) 条件概率, 这些条件概率可为标称特性值可信度的定量 (概率) 表述提供依据。附录 C.2 给出了其中两种表述方法。

B.2.4 除了简单的二元结果计数外, 针对具有多种可能结果的标称特性 (包括序值标度) 赋值的可靠性, 还可以使用一致性指标来表征。表 B.3 列出了此类指标的一些示例。相关参考文献中给出了公式。

注: 参考文献见: Eurachem/Citac Guide Assessment of the performance and uncertainty in qualitative chemical analysis. Edited by R. Bettencourt da Silva and S.L.R. Ellison 2021. ISBN 978-0-948926-39-6.

表 B.1 用于标称特性评估的计数实验

测试结果	已知或确认的测试样品状况		测试结果总数
	阳性	阴性	
阳性 (P)	n_{TP}	n_{FP}	$n_{P,obs}$
阴性 (P)	n_{FN}	n_{TN}	$n_{N,obs}$
总	n_P	n_N	N
注: 假定测试方法提供两种可能的结果, 阳性或阴性, 这些结果要么正确 (真), 要么错误 (假)。			

表 B.2 二元测试方法的性能参数

性能特性	公式 ^a	描述
假阳性率	$\frac{n_{FP}}{n_{TN} + n_{FP}} = \frac{n_{FP}}{n_N}$	阴性测试样品错误报告为阳性的比例，越低越好。
假阴性率	$\frac{n_{FN}}{n_{TP} + n_{FN}} = \frac{n_{FN}}{n_P}$	阳性测试样品错误报告为阴性的比例，越低越好。
灵敏度	$\frac{n_{TP}}{n_{TP} + n_{FN}} = \frac{n_{TP}}{n_P}$	阳性测试样品正确报告为阳性的比例，越高越好。
特异性	$\frac{n_{TN}}{n_{TN} + n_{FP}} = \frac{n_{TN}}{n_N}$	阴性测试样品正确报告为阴性的比例，越高越好。
阳性预测值	$\frac{n_{TP}}{n_{TP} + n_{FP}} = \frac{n_{TP}}{n_{P,obs}}$	阳性结果正确的比例，越高越好。
阴性预测值	$\frac{n_{TN}}{n_{TN} + n_{FN}} = \frac{n_{TN}}{n_{N,obs}}$	阴性结果正确的比例，越高越好。
效率	$\frac{n_{TP} + n_{TN}}{n_{TP} + n_{TN} + n_{FP} + n_{FN}}$	所有结果正确的比例，越高越好。
尤登指数	灵敏度(%) + 特异性(%) - 100	为方便起见，对性能数据的汇总，越高越好。
^a 公式提供了根据表 B.1 实验计数得到的基本比例或指数的估计值。		

表 B.3 定性测试方法性能参数示例

性能特性或程序	适用性			解释值的范围	参考文献
	赋值程序数量 ^a	测试样品数量 ^b	分类数量		
Cohen's kappa	2	□2	□2 ^c	-1 至+1 (越高越好)	[48] [49]
Fleiss's kappa	□2	□2	□2 ^c	-1 至+1 (越高越好)	[48] [50] [51]
属性一致性分析	□2	□2	□2 ^c	0%至 100% (越高越好)	ISO/TR 14468
Kendall 一致性系数	□3	□2	□2 (序值标尺)	0 至 1 (越高越好)	[49] [52]
分类方差分析 (CATANOVA)	□2	□2	□2	0 至 1 (越低越好) ^d	[28]
序值方差分析 (ORDANOVA)	□2	□2	□2 (序值标尺)	0 至 1 (越低越好) ^d	[29] [53]

^a 程序的数量可包括一组已知的参考值。也就是说，如果采用了单一标称值赋值程序，并与一组参考值进行对照检查，则赋值程序数量为 2。

^b 虽然大多数程序可以应用于仅有少量测试样品的情况，但随着样品数量的增加，结果会变得更具有信息量。

^c 尽管该程序适用于有多个可能响应的情况，但记录的结果只有一致或不一致两种，也就是说，评分是二元的。

^d Kendall 一致性系数使用在序数标尺上不同值之间的秩次差作为不一致性的量度。

注：与方差分析一样，ORDANOVA 提供了方差(不一致性)的定量量度；推荐的归一化量度范围为 0 到 1。值越高，表示一致性越低。解释上，采用基于模拟的显著性差异检验。

B.3 标称特性值的一致性或相似性定量指标

B.3.1 一致性指标简介

一些标称值是通过与参照对象进行大量特征上的比较来赋值。例如，通过 DNA 序列的实验测定和大量碱基对与特定物种参考序列的匹配，可实现生物物种的鉴定。在其他情况下，可以通过定性和定量观察结合的方法来识别某种材料。这时，观测值集与预期值集之间的一致性程度可以指示所赋特性值的可信度。

B.3.2 简要介绍了两种量度方法。

B.3.2 高沃（Gower）相似性指数

B.3.2.1 Gower 相似性指数，通常缩写为“Gower 指数”或“Gower 相似性”，已应用于 ISO/TR79 中介绍的定性特性标准物质的定值。它具有包容二元、序值标尺和定量信息的优点。其计算方法见公式（B.1）和（B.2）：

$$g = \frac{\sum_{i=1}^n w_i g_i}{\sum_{i=1}^n w_i} \quad (\text{B.1})$$

$$g_i = 1 - \frac{|\square\square - \square_{0,\square}|}{r_i} \quad (\text{B.2})$$

$$g = \frac{\sum_{i=1}^n w_i g_i}{\sum_{i=1}^n w_i} \quad (\text{B.1})$$

$$g_i = 1 - \frac{|x_i - x_{0,i}|}{r_i} \quad (\text{B.2})$$

其中：g 为根据公式(B.1)定义的 Gower 相似系数；

g_i 为第 i 个特征的相似度量度；

w_i 为分配给第 i 个特征的权重；

x_i 为第 i 个特征的观测值；

$x_{0,i}$ 为第 i 个特征的选定参考值；

r_i 为第 i 个特征的取值范围。

B.3.2.2 范围 r_i 的选择使得 g_i 介于 0 和 1 之间。权重 w_i 根据实际情况进行赋值。 g 值接近 1 表明与预期值高度一致； g 值接近 0 表明一致性差。该指数可用作标称特性的赋值标准，或用于在报告或标准物质文件中表达匹配的一致程度。

注：针对标准物质，Gower 指数的计算和应用详见 ISO/TR 79。

B.3.3 核酸或蛋白质序列匹配评分

生物物种常根据 DNA 或蛋白质序列的匹配来进行鉴定或确认。对于核酸或蛋白质序列，序列相似度用来表示两个序列之间的相似程度。通常，相似度通过评分来量化，并有多评分系统可供使用。相似度一般采用与相同或更高评分的随机匹配概率来表示。该方法可在谨慎使用的前提下（见 7.3.3 和 7.3），提供对物种赋值可信度的定量表达。

注 1：具体评分系统参考以下文献：] Karlin, S. and ALTSCUL, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proc Natl Acad Sci U S A. 1990, 87(6), 2264-2268. doi:10.1073/pnas.87.6.2264。

注 2：序列相似性可以表示为“一致性百分比”，即两个序列之间特征匹配的百分比。

B.4 标称特性值可信度的其他常用表达

B.4.1 似然比

B.4.1.1 似然比可以用来传达由特定测试结果提供的“证据权重”。似然比提供了一种假设比另一种假设更具胜算的可能性增量，该增量来自于新的信息，比如如一个阳性的测试结果。对于标称特性值，所关注的假设是 H_1 ：所赋值是正确的， H_2 ：所赋值是不正确的。假设对候选标准物质标称特性的测试结果为阳性，所赋值相应也为阳性，似然比 $L_{H_1:H_2}$ 由公式 (B.3) 给出。

$$\square_{H_1:H_2} = \frac{1 - \square(\square|\square)}{\square(\square|\square)} = \frac{1 - \square(\square|\square)}{\square(\square|\square)} \quad (B.3)$$

其中：

$p(N|P)$ 为表 B.1 中假阴性率的估计值(这里写为给定阳性测试样品的阴性响应的条件概率)；

$p(P|N)$ 为相应的假阳性率，再次写为条件概率的形式；

$p(P|P) = 1 - p(N|P)$ ，为对给定阳性测试样品给出阳性结果的概率，即真阳性率。

对于该简单情况，似然比为真阳性率除以假阳性率。

B.4.1.2 公式 (B.3) 给出的似然比可以取任意正值。值大于 1 表示支持假设 H_1 ；值小于 1 表示支持假设 H_2 。对于标准物质赋值有关的测试，假阳性和假阴性的误判率通常都很低。因此，似然比的预期通常超过 1000。当采用多组独立测试时，似然比可超过 10^9 。

注：似然比高并不能保证结果正确，也不能用于表达测试结果正确的概率。它体现的测试结果支持赋值的程度是相对于相反赋值的。高似然比提供了支持报告值的强有力证据，但如果信息表明似然比不适用，则不应采用。

B.4.2 后验概率

B.4.2.1 后验概率是贝叶斯统计中的一个概念。它可以用于直接表示所赋标称特性值为正确的概率。

B.4.2.2 后验概率结合了候选标准物质特定标称特性值的先验信息和额外辅助信息，该辅助信息通常用于增加最终计算出的标称特性值的概率。这一过程的基础是贝叶斯定理，对于标准物质的赋值，可用公式 (B.4) 表达。

$$p(A|\square) = p(\square) \frac{p(\square|\square)}{p(\square)} \quad (\text{B.4})$$

其中：

$p(A|E)$ 为由证据 E (通常是一个或多个测试结果) 得到赋值 A 的概率，被称作 A 的后验概率；

$p(A)$ 为赋值前与赋值 A 有关的概率，即 A 的先验概率；

$p(E|A)$ 为由标称特性值为 A 的标准物质，得到证据(或测试结果) E 的概率；

$p(E)$ 为获得测试结果 E 的概率，与标称特性的值无关。

注 1： $p(x|y)$ 形式的条件概率通常被解读为“给定事件 y ，事件 x 的概率”，或简略地称为“给定 y 的 x 的概率”。

注 2： $p(E)$ 是针对所研究的总体进行评估的，该总体中包含比例可高可低的、具有特性 A 的样品。有时写作 $p(E) = p(E|A) + p(E|\bar{A})$ ，其中 $\bar{A}$ 表示“非 A ”，以清晰反映对样品值不为 A 但测试结果为 A ，即假阳性结果的关注。

注 3： 公式 (B.4) 中的分数 $p(E|A)/p(E)$ 有时被称作似然比，因此，后验概率为先验概

率乘以似然比。

公式 (B.4) 原则上可以递归式应用于一系列独立测试结果, 将前一次计算的后验概率作为下一次的先验概率, 理论上, 公式 (B.4) 支持对一系列独立测试结果进行递归运算。具体做法是, 将上一步计算出的后验概率作为下一步的先验概率输入。依此类推, 即可在汇总多次独立测试信息后, 得到最终的后验概率。

公式 (B.4) 的应用也存在不少潜在缺陷, 包括但不限于以下方面:

- 1) 很难确定可靠的先验概率 $p(A)$;
- 2) 很难获得给定总体的可靠概率 $p(E)$;
- 3) 难以定义所研究的总体以及该总体中正确或错误地产生测试结果 E 的比例;
- 4) 对测试结果或观测值之间缺少独立性非常敏感。例如, 在光谱学中, 单独的信号很少完全相互独立, 而是可以在相关化合物中同时出现。

B.4.2.4 B.4.2.3 中列出的潜在缺陷可能导致后验概率的估算偏差达到几个数量级。因此, 若想成功应用公式 (B.4), 需要具备如下条件:

- 1) 关于所研究样品总体的可靠信息;
- 2) 针对总体不同部分的假阳性率与假阴性率的准确信息;
- 3) 经确认的、针对所有相关测试的错误响应率;
- 4) 不同测试结果之间任何关联性 (非独立性) 的完整信息;
- 5) 具备应用公式 (B.4) 所需的有适当资质和经验的人员。

B.4.3 熵

B.4.3.1 对于标称特性赋值的不确定度, 最充分的表达方式是该特性可能取值的概率分布。

例: 在确定 DNA 链中特定位点的碱基时, 测序仪可能得出以下概率分布: 分布 a (A、C、G、T 的概率分别 0.04、0.09、0.06、0.81), 与分布 b (A、C、G、T 的概率分别 0.02、0.00、0.01、0.97) 相比, 不确定度更大。

B.4.3.2 熵用于衡量概率分布在特性所有可能取值集合上的分散程度。如果该集合为 $\{a, b, c, \dots\}$, 相应的概率为 $\{p_a, p_b, p_c, \dots\}$ (非负数, 和为 1), 则该概率分布的熵为 $H = -(p_a \times \ln(p_a) + p_b \times \ln(p_b) + p_c \times \ln(p_c) + \dots)$, 其中“ln”表示自然对数 (底数 $e=2.718282\dots$)。由于 H 越大, 不确定度越大, 因此熵可以用于表达标称特性赋值的不确定度。

注：可参考以下文献中的示例 E6: Possolo A. Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results, NIST Technical Note 1900, 2015 National Institute of Standards and Technology, Gaithersburg, MD. doi: 10.6028/NIST.TN.1900。

例：对于 B.4.3.1 中的分布 a, $H=-(0.04 \times \ln(0.04) + 0.09 \times \ln(0.09) + 0.06 \times \ln(0.06) + 0.81 \times \ln(0.81)) = 0.685$ ；对分布 b, $H=$

$-(0.02 \times \ln(0.02) + 0.00 \times \ln(0.00) + 0.01 \times \ln(0.01) + 0.97 \times \ln(0.97)) = 0.154$ ，其中： $0 \times \ln(0) = 0$ 。

B.4.3.3 具有概率密度函数 p 的连续概率分布的熵是 $-p \times \ln(p)$ 在实数集上的积分，其中 $p > 0$ 。对于均值为 μ 、标准差为 σ 的高斯（或正态）分布，熵为 $H = \ln(\sigma) + 0.5 \times \ln(2\pi e)$ ，因此， σ 与 $\exp(H)$ 成正比。尽管这只是一种特殊情况，但它揭示了 H 增大时 σ 也随之增大，并暗示 $\exp(H)$ 通常可被视作标准不确定度的一种对应量（相差一个比例因子）。

例：在某法医调查中，基于某玻璃碎片的折射率和主要氧化物的质量分数，分别得出其来源的以下概率：0.36（现代建筑窗户）、0.56（老旧建筑窗户）、0.08（汽车窗户）、0.00（瓶子）、0.00（餐具）、0.00（汽车前灯）。该分布的熵为 $-(0.36 \times \ln(0.36) + 0.56 \times \ln(0.56) + 0.08 \times \ln(0.08)) = 0.895$ ，其“等效”的标准不确定度为 $\exp(0.895) / \exp(0.5 \times \ln(2\pi e)) = 0.592$ 。

附录 C

常见统计学程序介绍（资料性）

C.1 二维列联表的卡方检验

C.1.1 目的

以下程序提供了一种对不同结果组中不合格结果比例的显著性差异进行检验的方法，可用于：

- 1) 在均匀性研究，即对每个标准物质单元进行重复测试后，检验各单元之间是否存在显著性差异（见第 9 章）；
- 2) 在稳定性研究，即将多个标准物质单元暴露于不同条件下并进行后续测试后，检验是否存在时间效应或其他条件（如温度）影响的证据（见第 10 章）。

注：此处所述的程序适用于单一分组因素，该因素可以是标准物质单元（在均匀性研究的情况下）或单一处理条件，如时间或温度（在稳定性研究的情况下）。

C.1.2 数据要求

该程序需要以下数据：

- 1) 对于均匀性研究：标准物质单元的合格与不合格（或阴性与阳性）结果的计数表。
- 2) 对于稳定性研究：按暴露条件列出的未失效与失效（或阴性与阳性）结果的计数表。

表格样式示例见表 C.1 和表 C.2。

注：为便于排版，本规范中的表格均以横向布局呈现。对于卡方检验，表格的行列方向不影响计算结果，可视情况对调行和列。

C.1.3 程序

记表的列数为 n_c ，行数为 n_R 。将表中第 i 行第 j 列的单元格中的计数记为 n_{ij} 。然后进行以下计算：

- 1) 根据公式 (C.1) 计算每行总的计数 r_i ：

$$r_i = \sum_{j=1}^{n_c} n_{ij} \quad (C.1)$$

2) 根据公式(C.2)计算每列总的计数 c_j :

$$c_j = \sum_{i=1}^{n_R} n_{ij} \quad (\text{C.2})$$

3) 根据公式(C.3)计算总观测数 N :

$$N = \sum_{i=1}^{n_R} \sum_{j=1}^{n_C} n_{ij} \quad (\text{C.3})$$

4) 根据公式 (C.4) 计算表格中每个单元格的期望计数 E_{ij} :

$$E_{ij} = \frac{r_i \cdot c_j}{N} \quad (\text{C.4})$$

4) 根据公式(C.5)计算卡方检验统计量 χ^2 。

$$\chi^2 = \sum_{i=1}^{n_R} \sum_{j=1}^{n_C} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \quad (\text{C.5})$$

5) 根据公式(C.6)计算自由度 ν 。

$$\nu = (n_R - 1)(n_C - 1) \quad (\text{C.6})$$

7) 从卡方分布的分位数表或软件中获得临界值 $\chi^2_{0.95, \nu}$, 即自由度为 ν 的卡方分布的上 0.95 分位数。

8) 将检验统计量 χ^2 与临界值 $\chi^2_{0.95, \nu}$ 进行比较, 如果检验统计量 χ^2 大于临界值 $\chi^2_{0.95, \nu}$, 则表明检验结果在 95%置信水平下显著。

对于均匀性检验, 检验结果为显著, 表明材料不均匀。对于稳定性测试, 检验结果为显著, 则表明不同时间的失效单元比例存在差异。

注: 当表中所有单元格的计数均不小于 5, 则 C.1.3 中第 5 项定义的检验统计量的参考概率分布近似为卡方分布。若较多单元格的计数小于 5, 则可改用费舍尔精确检验

(Fisher's exact test), 参考文献见: Agresti A. Categorical Data Analysis, John Wiley & Sons, Hoboken, NJ, 3rd edition, 2013. ISBN 978-0-470-46363-5。

注: 利用卡方检验解释稳定性测试结果见 10.2.3。

表 C.1 基本均匀性研究表格样式示例

测试 结果	标准物质单元号						
	#1	#2	#3	...	#(j)	...	# (n_p)
合格	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	...	$n_{1,j}$	...	n_{1,n_C}
不合格	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	...	$n_{2,j}$	...	n_{2,n_C}
注: n_{ij} 代表每个单元格中的观测值 (分别为合格或不合格) 数量。							

表 C.2 基本稳定性研究表格样式示例

测试 结果	时间点					
	#1	#2	...	#(j)	...	# (n _p)
未失效	n _{1,1}	n _{1,2}	...	n _{1,j}	...	n _{1,n_c}
失效	n _{2,1}	n _{2,2}	...	n _{2,j}	...	n _{2,n_c}

注：n_{ij}代表每个单元格中的观测值（分别为未失效或失效）数量。

C.2 标准物质单元不合格/合格率的置信限

C.2.1 标准物质单元不合格率的上置信限

标准物质单元不合格率 $\hat{p}_{NC}$ 的 95% 上置信限可基于威尔逊 (wilson) 评分区间计算, 并需要以下数据信息:

- 1) 均匀性研究中测试的标准物质单元数量 n_h ;
- 2) 研究中发现的不合格标准物质单元数量 n_{NC} ;
- 3) 要求的置信水平 $1-a$, 对于 95% 上限, $a=0.05$;
- 4) 正态分布的 $(1-a)$ 分位数 z_{1-a} , 对于 95% 上限, $z_{1-a}=1.645$ 。

计算程序如下:

- 1) 根据公式 C.7 计算估计的不合格率 $\hat{p}_{NC}$:

$$\hat{p}_{NC} = \frac{n_{NC}}{n_h} \quad (C.7)$$

- 2) 根据公式 C.8 计算不合格率的 95% 上置信限 $\hat{L}_{p_{NC}, u}$ 。

$$\hat{L}_{p_{NC}, u} = \frac{\hat{p}_{NC} + \frac{z_{1-a}^2}{2n_h} + \frac{z_{1-a}}{1 + \frac{z_{1-a}^2}{n_h}} \sqrt{\frac{\hat{p}_{NC}(1-\hat{p}_{NC})}{n_h} + \frac{z_{1-a}^2}{4n_h}}}{1 + \frac{z_{1-a}^2}{n_h}} \quad (C.8)$$

注: 对于其他置信水平下置信限的计算, 可将 z_{1-a} 替换为对应置信水平下的正态分布分位数。

例: 在均匀性测试中, 检查了 120 个标准物质单元, 其中发现有 2 个单元的响应不符合预期, 被标记为不合格。则:

$$\hat{p}_{NC} = \frac{2}{120} = 0.0167$$

$$\hat{L}_{p_{NC}, u} = \frac{0.0167 + \frac{1.645^2}{2 \times 120} + 1.645 \sqrt{\frac{0.0167(1-0.0167)}{120} + \frac{1.645^2}{4 \times 120^2}}}{1 + \frac{1.645^2}{120}} = 0.049$$

C.2.2 标准物质单元合格率的下置信限

公式 C.9 用于计算标准物质单元合格率的 95%下置信限 $\hat{L}_{p_C, l}$ ：

$$L_{p_{C,l}} = 1 - L_{p_{NC}, u} \quad (C.9)$$

$L_{p_{C,l}}$ ：可报告为比率或百分比。

例：若计算得到的标准物质单元不合格率的 95%上置信限为 0.049，则标准物质单元合格率的 95%下置信限为 $1-0.049=0.951$ 或 95.1%，可报告为“合格率不低于 95%”。

全国标准物质计量技术委员会征求意见稿

附录 D

定性特性标准物质的均匀性和稳定性研究示例（资料性）

D.1 肉类掺假识别标准物质的均匀性研究

D.1.1 概述

候选标准物质的预期用途是为检测牛肉中是否存在猪肉提供控制物质。该候选标准物质通过向牛肉馅中加入标称 5% 的猪肉馅制备而成。该示例用于说明在定性测试方法有显著的假阴性率的情况下，如何进行均匀性的卡方检验。此外，还计算了不合格单元比例的上置信限。

D.1.2 研究设计和结果

从一批约 200 个包装好的候选标准物质单元中随机选择 18 个单元，并使用商业猪肉蛋白免疫试剂盒对每个单元进行 5 次测试。如果试剂盒颜色强度超过规定阈值，则试剂盒测试结果为阳性。在两个免疫分析测试板上以完全随机的顺序进行重复测试。结果汇总见表 D.1。

注 1：为便于说明，试剂盒的阈值被提高到正常操作阈值以上，以对已知阳性测试样本有大约 10% 的错误率。

注 2：此设计对应于 JJF1343 中的均匀性实验“基本”设计。

表 D.1 候选阳性定性标准物质的均匀性测试结果

测试结果	标准物质单元号																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
阳性 ^a	4	4	5	5	4	5	5	5	5	5	5	5	5	4	5	4	3	3
阴性 ^a	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	2	2
^a 表项中的输入的是每个标准物质单元的阳性和阴性结果数量。																		

D.1.3 均匀性的卡方检验

可按照 C.1 中的程序进行卡方检验：

1) 计算每行的总数 r_i 。将阳性结果的行总数记为 r_p ，阴性结果的行总数记为 r_N ，通过对每行的所有单元格求和得到：

$$r_P = 4 + 4 + 5 + \cdots + 4 + 3 + 3 = 81 \quad (D.1)$$

$$r_N = 1 + 1 + 0 + \cdots + 1 + 2 + 2 = 9 \quad (D.2)$$

其中"..."表示省略的中间单元格。

2) 计算每列的总 c_j 。对于表 D.1, 由于每个标准物质单元都有 5 次重复测试且没有缺失观测值, 因此所有列的总和为 5, 所以:

$$\text{对于 } j=1, 2, \dots, 18, c_j=5 \quad (D.3)$$

3) 计算总观察次数 N :

$$N = \sum_{i=1}^{n_R} \sum_{j=1}^{n_C} n_{ij} = 90 \quad (D.6)$$

4) 计算每个单元格的期望计数。由于所有列的观察次数相同, 因此“阳性”行中所有单元格的期望计数是相同的, “阴性”行中所有计数也是相同的。将“阳性”行和“阴性”行的期望计数分别记为 E_P 和 E_N , 并将所有列的计数设为恒定 $c_{all}=5$, 见 2):

$$E_P = \frac{r_P \cdot c_{all}}{N} = \frac{81 \times 5}{90} = 4.5 \quad (D.5)$$

$$E_N = \frac{r_N \cdot c_{all}}{N} = \frac{9 \times 5}{90} = 0.5 \quad (D.6)$$

5) 根据公式 C.5 计算卡方检验统计量 χ^2 。这涉及对如下形式的项求和:

$$\chi^2 = \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \quad (D.7)$$

例如, 对于第 1 行第 1 个标准物质单元(阳性行)的单元格, $n_{ij}=4$, $E_{ij}=E_P=4.5$, 因此对 χ^2 的贡献为 $(4 - 4.5)^2 / 4.5 = 0.0556$ (精确到小数点后四位)。对所有这样的项求和得到: $\chi^2=18.89$ 。

6) 计算自由度 ν :

$$\nu = (n_R - 1)(n_C - 1) = (2 - 1)(18 - 1) = 17 \quad (D.8)$$

7) 得到临界值 $\chi^2_{0.95, \nu}$ 。对于自由度 $\nu=17$, 临界值 $\chi^2_{0.95, 17} = 27.58$ 。

8) 将检验统计量 χ^2 与临界值进行比较。在该情况下, 检验统计量为 18.89, 而临界值为 27.58。检验结果在 95% 的置信水平下不显著, 因此没有标准物质单元间不均匀性的证据。

D.1.4 标准物质单元不合格率的置信限

采用公式 C.8 计算标准物质不合格单元率的 95% 上置信限。根据 9.3.2.2, 需要对不合格单元进行计数。当测试方法有显著的错误响应率时, 判断这一点可能会很困难。考虑到测试方法可能会对合格测试样本(这里指阳性样本)产生约 10%

的错误响应，若有 3 次以上的阴性响应，则判定样本为不合格。在这种情况下，结果为 18 个合格标准物质单元和 0 (零) 个不合格单元。对于公式 C.8， $n=18$ ， $n_{NC}=0$ ， $\hat{p}_{NC}=0/18=0.0$ 。因此不合格单元率的上限 $\hat{L}_{p_{NC}, u}$ 为：

$$L_{p_{NC}, u} = \frac{0.0 + \frac{1.645^2}{2 \times 18} + 1.645 \sqrt{\frac{0.0(1-0.0)}{18} + \frac{1.645^2}{4 \times 18^2}}}{1 + \frac{1.645^2}{18}} = 0.13$$

根据 C.2.2，也可以报告为标准物质单元合格率不低于 $1-0.13=0.87$ ，即不低于 87%。

D.2 加速短期（运输）稳定性研究

D.2.1 概述

候选标准物质中含有标称质量分数为 1% 的马肉，预期用途是为检测牛肉中是否存在马肉提供控制物质。该研究考察了该物质在不同温度下的稳定性，时间为一个月。示例采用简单的卡方检验来判断时间和/或温度对马肉蛋白试剂盒定性响应的显著性影响。

D.2.2 研究设计和结果

该研究考察了 -20°C、4°C、18°C、37°C 和 60°C 温度下的暴露情况。分别有 3 个标准物质单元在每个温度和或时间组合下保存 0、2 或 4 周，时间点 0 对应于瞬时暴露并返回 -20°C 下保存。将所有标准物质单元随机分配到不同的温度和时间条件下。暴露结束后返回 -20°C 下保存，并采用商用的酶联免疫吸附 (ELISA) 试剂盒进行同时分析，形成同步设计。每组条件下检测到马肉阳性（检出）或阴性（未检出）的单元数量见表 D.2。

D.2.3 检查

通过检查表 D.2 中的结果可以看出，在温度高达 37°C 时，没有失效单元，而所有暴露于 60°C，超过瞬时暴露时间（0 周）的单元均失效。即使不进行统计分析，通常也足以将其作为标准物质运输温度保持 60°C 以下，理想情况 37°C 以下的充分理由。但是，统计学分析可用于确认所观察到的暴露条件与失效率之间的关联并非偶然。

D.2.4 稳定性显著影响的卡方检验

D.2.4.1 稳定性研究包括多个时间点和多个温度，相对表 C.2 中仅按时间点对阳性和阴性结果分组的简单研究布局更为复杂。但是，可通过将每个不同时间和温

度的组合视为一个独立的处理，从而简化为双向列联表。表 D.2 说明了这一点，即每一行对应于不同的处理，列对应于不同的观测结果。表 D.2 的两个“结果”列因此构成了一个双向列联表，可按照 C.1.3 的程序进行分析。

注：将双向列联表中的行和列调换，不会影响卡方检验的结果。

D.2.4.2 基于表 D.2 的双向布局以及表中显示的行、列合计，卡方检验所需的信息如下：

- 1) 观测总数 $N=45$ 。
- 2) 列总数：阳性（检出）为 39；阴性（未检出）为 6。
- 3) 行总数：每行 3 个（所有行相同）。
- 4) 期望值：如前所述，由于所有处理的观测总数相同，只需分别计算所有阳性（检出）和所有阴性（未检出）单元格的期望值 E_P 和 E_N 。行计数均为 3，根据表 D.2 中的列总数，得到 $E_P=3 \times 39/45=2.6$ ； $E_N=3 \times 6/45=0.4$ 。
- 5) 自由度 $\nu=(15-1)(2-1)=14$ 。
- 6) 临界值 $\chi^2_{0.95,14}=23.68$ 。

表 D.2 牛肉中马肉稳定性研究结果汇总（采用 ELISA 方法检测马肉）

条件			结果（单元数）		行合计
处理	温度（℃）	时间（周）	检出（未失效）	未检出（失效）	
A	-20	0	3	0	3
B	-20	2	3	0	33
C	-20	4	3	0	3
D	4	0	3	0	3
E	4	2	3	0	3
F	4	4	3	0	3
G	18	0	3	0	3
H	18	2	3	0	3
I	18	4	3	0	3
J	37	0	3	0	3
K	37	2	3	0	3
L	37	4	3	0	3
M	60	0	3	0	3
N	60	2	0	3	3

P	60	4	0	3	3
合计			39	6	

D.2.4.3 卡方检验统计量的计算涉及列联表中所有单元格贡献的求和。D.1.3 中给出了 χ^2 检验统计量的贡献项的示例。考虑处理 H (18°C, 2 周) 的“检出”列, 表单元格的计数 n_{ij} 为 3。这是阳性 (检出) 响应的计数, 因此适用的期望值是 E_P 。对总卡方检验统计量的贡献为 $(3-2.6)^2/2.6=0.0616$ (精确到小数点后四位)。其他贡献采用表单元格中相应的计数和期望值, 以相同方式计算。所有这些贡献的总合为 $\chi^2=45$ 。

D.2.4.4 将计算得到的检验统计量 (45) 与临界值 (23.68) 进行比较, 确认检验统计量远高于临界值。因此, 卡方检验结果在 95% 的置信水平上显著。

注 1: 使用软件来确定与自由度 14 的卡方检验统计量 45 相关的概率, 得到的 p 值约为 4×10^{-5} , 这可作为强有力的证据, 表明处理与 RM 单元失效之间的关联绝非偶然。

注 2: 将结果简化为双向表虽然方便, 但无法显示哪个实验因素是最重要的。采用图形检查方法 (例如按不同因素分组的失效比例条形图), 有助于根据卡方检验显著结果进一步识别出最重要的因素。