CNAS—GL002

能力验证结果的统计处理和能力评价指南

Guidance on Statistic Treatment of Proficiency Testing Results and Performance Evaluation

中国合格评定国家认可委员会

2018 年 3 月 1 日 发布　　2018 年 3 月 1 日 实施

目 次

前 言

本文件为能力验证结果的统计处理和能力评价提供指南。

本文件依据 GB/T 27043《合格评定 能力验证的通用要求》制订，同时参考了 GB/T 28043《利用实验室间比对进行能力验证的统计方法》。GB/T 28043 给出了能力验证统计方法的更详细指南，使用本文件时，可同时参考 GB/T 28043。

能力验证结果的统计处理和能力评价指南

1 范围

1.1 本文件为能力验证结果的统计处理和能力评价提供指南。

1.2 本文件适用于 CNAS 的能力验证，也可为其他机构组织能力验证提供参考。

2 规范性引用文件

下列文件中的条款通过引用而成为本文件的条款。以下引用的文件，注明日期的，仅引用的版本适用；未注明日期的，引用文件的最新版本（包括任何修订）适用。

CNAS-RL02 能力验证规则

CNAS-GL03 能力验证样品均匀性和稳定性评价指南

GB/T 27043 合格评定 能力验证的通用要求（ISO/IEC 17043，IDT）

GB/T 28043 利用实验室间比对进行能力验证的统计方法（ISO 13528，IDT）

GB/T 6379 测量方法与结果的准确度（正确度和精密度）（ISO 5725，IDT）

ISO/IEC 指南 98-3 测量不确定度 第 3 部分：测量不确定度的表示指南

ISO/IEC 指南 99：2007 国际计量学词汇 基础和通用概念及相关术语

IUPAC 技术报告 分析化学实验室能力验证国际协议

3 术语和定义

CNAS-RL02、GB/T 27043、GB/T 28043、ISO/IEC 指南 99 界定的术语和定义适用于本文件。为方便使用，重复列出以下术语和定义：

3.1 实验室间比对 interlaboratory comparison

按照预先规定的条件，由两个或多个实验室对相同或类似的物品进行测量或检测的组织、实施和评价。

3.2 能力验证 proficiency testing

利用实验室间比对，按照预先制定的准则评价参加者的能力。

3.3 指定值 assigned value

对能力验证物品的特定性质赋予的值。

3.4 能力评定标准差 standard deviation for proficiency assessment

根据可获得的信息，用于评价能力验证结果分散性的度量。

注 1：标准差只适用于比例尺度和定距尺度的结果。

注 2：并非所有的能力验证计划都根据结果的分散性进行评价。

3.5 z 比分数 z-score

由能力验证的指定值和能力评定标准差计算的实验室偏倚的标准化度量。

注：z 比分数有时也称为 z 值或 z 分数。

3.6 离群值 outlier

一组数据中被认为与该组其他数据不一致的观测值。

注：离群值可能来源于不同的总体，或由于不正确的记录或其他粗大误差的结果。

3.7 稳健统计方法 robust statistical method

对给定概率模型假定条件的微小偏离不敏感的统计方法。

3.8 测量审核 measurement audit

一个参加者对被测物品(材料或制品)进行实际测试，其测试结果与参考值进行比较的活动。

注：测量审核是对一个参加者进行“一对一”能力评价的能力验证计划。

4 统计处理和能力评价

4.1 总则

能力验证的结果可以以多种形式出现，并构成各种统计分布。分析数据的统计方法应与数据类型及其统计分布特性相适应。分析这些结果时，应根据不同情况选择适用的统计方法。各种情况下优先使用的具体方法，可参见 GB/T 28043。对于其他方法，只要具有统计依据并向参加者进行了详细描述，也可使用。无论使用哪一种方法对参加者的结果进行评价，一般包括以下几方面内容：

a）指定值的确定；

b）能力统计量的计算；

c）能力评定。

必要时，考虑能力验证物品的均匀性和稳定性对能力评定的影响。能力验证物品均匀性和稳定性的评价方法见 CNAS-GL003《能力验证样品均匀性和稳定性评价指南》、GB/T 28043 和 IUPAC 技术报告。

4.2 统计设计

4.2.1 应根据数据的特性（定量或定性，包括顺序和分类）、统计假设、误差的性质以及预期的结果数量，制定符合计划目标的统计设计。在统计设计中应考虑下列事项：

a）能力验证中每个被测量或特性所要求或期望的准确度（正确度和精密度）以及测量不确定度；

b）达到统计设计目标所需的最少参加者数量；当参加者数量不足以达到目标或不能对结果进行有意义的统计分析时，应将评定参加者能力的替代方法的详细内容提供给参加者；

c）有效数字与所报告结果的相关性，包括小数位数；

d）需要检测或测量的能力验证物品数量，以及对每个能力验证物品或每项测定的检测、校准或测量的重复次数；

e）用于确定能力评定标准差或其它评定准则的程序；

f）用于识别和（或）处理离群值的程序；

g）只要适用，对统计分析中剔除值的评价程序；

h）只要适当，与设计相符的目标和能力验证轮次的频率。

4.2.2 在缺少统计设计所需的可靠信息时，可通过开展先期实验室间比对来获得。

4.3 指定值及其不确定度的确定

4.3.1 指定值的确定有多种方法，以下列出最常用的方法。在大多数情况下，按照以下次序，指定值的不确定度逐渐增大。

a）已知值 —— 根据特定能力验证物品配方（如制造或稀释）确定的结果；

b）有证参考值 —— 根据定义的检测或测量方法确定（针对定量检测）；

c）参考值 —— 根据对能力验证物品和可溯源到国家标准或国际标准的标准物质/标准样品或参考标准的并行分析、测量或比对来确定；

d）由专家参加者确定的公议值 —— 专家参加者（某些情况下可能是参考实验室）应当具有可证实的测定被测量的能力，并使用已确认的、有较高准确度的方法，且该方法与常用方法有可比性；

e）由参加者确定的公议值 —— 使用 GB/T 28043 和 IUPAC 国际协议等给出的统计方法，并考虑离群值的影响。例如，以参加者结果的稳健平均值、中位值（也称为中位数）等作为指定值。附录 A 给出了由参加者结果确定指定值的常用稳健统计方法。

4.3.2 对上述每类指定值的不确定度，可参照 GB/T 28043 等所描述的方法进行评定。此外，ISO/IEC 指南 98-3 中给出了确定不确定度的其它信息。

4.3.3 指定值的确定应确保公平地评价参加者，并尽量使检测或测量方法间吻合一致。只要可能，应通过选择共同的比对小组以及使用共同的指定值达到这一目的。

4.3.4 对定性数据[也称为“分类的”或“定名的”值]或半定量值[也称为“顺序的”值]，其指定值通常需要由专家进行判断或由制造过程确定。某些情况下，可使用大多数参加者的结果（预先确定的比例，如 80%或更高）来确定公议值。该比例应基于能力验证计划的目标和参加者的能力和经验水平来确定。

4.3.5 离群值可按下列方法进行统计处理：

a）明显错误的结果，如单位错误、小数点错误、计算错误或者错报为其他能力验证物品的结果，应从数据集中剔除，单独处理。这些结果不再计入离群值检验或稳健统计分析。明显错误的结果应由专家进行识别和判断。

b）当使用参加者的结果确定指定值时，应使用适当的统计方法使离群值的影响降到最低，即可以使用稳健统计方法或计算前剔除离群值。

c）如果某结果作为离群值被剔除，则仅在计算总计统计量时剔除该值。但这些结果仍应当在能力验证计划中予以评价，并进行适当能力评定。

4.3.6 需考虑的其他事项

a）理想情况下，如果指定值由参加者公议确定，应当有确定该指定值正确度和检查数据分布的程序。例如，可采用将指定值与一个具备专业能力的实验室得到的参考值进行比较等方法确定指定值的正确度。

通常，正态分布是许多数据统计处理的基础。正态分布的特点是单峰性、对称性、有界性和抵偿性。作为一个能力验证计划的结果，由于参加者的测试方法、测试条件往往各不相同，而且能力验证结果的数量也是有限的，所以在许多情况下能力验证的结果呈偏态分布。对能力验证的结果只要求近似正态分布，尽可能对称，但分布应当是单峰的，如果分布中出现双峰或多峰，则表明参加者之间存在群体性的系统偏差，这时应研究其原因，并采取相应的措施。例如，可能是由于使用了产生不同结果的两种检测方法造成的双峰分布。在这种情况下，应对两种方法的数据进行分离，然后对每一种方法的数据分别进行统计分析。数据直方图或核(Kernel)密度图可以显示结果的分布情况。

b)应当有依据不确定度来判断指定值是否可接受的准则。在 GB/T 28043 和 IUPAC 国际协议中给出了该准则，该准则是基于限定指定值不确定度对能力评定的影响而建立的，即：准则限定了由于指定值的不确定度而使参加者得到一个不可接受的评估结果的可能性。

4.4 能力统计量的计算

4.4.1 定量结果

4.4.1.1 能力验证结果通常需要转化为能力统计量，以便进行解释和与其他确定的目标作比较。其目的是依据能力评定准则来度量与指定值的偏离。所用统计方法可能从不做任何处理到使用复杂的统计变换。

注：“能力统计量”也称为“性能统计量”。

4.4.1.2 能力统计量对参加者应是有意义的。因此，统计量应适合于相关检测，并在某特定领域得到认同或被视为惯例。

4.4.1.3 按照对参加者结果转化由简至繁的顺序,定量结果的常用统计量如下：

a）差值D，由（1）式计算：

$$D = x - X \qquad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

式中：

x为参加者结果；

X 为指定值。

b）百分相对差$D_{\%}$，由（2）式计算：

$$D_{\%} = \frac{(x - X)}{X} \times 100 \qquad \cdots\cdots\cdots\cdots\cdots\cdots (2)$$

c）z比分数，由（3）式计算：

$$z=\frac{x-X}{\hat{\sigma}} \quad \cdots\cdots\cdots\cdots\cdots\cdots (3)$$

式中：

$\hat{\sigma}$ 为能力评定标准差。$\hat{\sigma}$ 可由以下方法确定：

—— 与能力评价的目标和目的相符，由专家判定或法规规定（规定值）；

—— 根据以前轮次的能力验证得到的估计值或由经验得到的预期值（经验值）；

—— 由统计模型得到的估计值（一般模型）；

—— 由精密度试验得到的结果；

—— 由参加者结果得到的稳健标准差、标准化四分位距、传统标准差等。

具体方法参见附录 A 和 GB/T 28043 等。

d）z'比分数，由式（4）计算：

$$z'=(x-X)/\sqrt{\hat{\sigma}^2+u_X^2} \quad \cdots\cdots\cdots\cdots\cdots\cdots (4)$$

式中：

u_X 为指定值的标准不确定度。

注 1：当指定值的确定未用到参加者的结果时，可用式（4）来计算。

注 2：z'比分数有时也称作 z'分数或 z'值。

e）ζ 比分数，由式（5）计算，除了使用标准不确定度代替扩展不确定度外，计算与 E_n 值类似。

$$\zeta=\frac{x-X}{\sqrt{u_x^2+u_X^2}} \quad \cdots\cdots\cdots\cdots\cdots\cdots (5)$$

式中：

u_x 为参加者结果的合成标准不确定度。

注 1：仅当 x 和 X 不相关时，式（5）才成立。

注 2：ζ 比分数有时也称作 ζ 分数或 ζ 值。

f）E_n 值，由式（6）计算：

$$E_n=\frac{x-X}{\sqrt{U_x^2+U_X^2}} \quad \cdots\cdots\cdots\cdots\cdots\cdots (6)$$

式中：

U_x 为参加者结果的扩展不确定度；

U_X 为指定值的扩展不确定度；

U_x 和 U_X 的包含因子 k=2。

注 1：E_n 值有时也称作 E_n 数。

注 2：仅当 x 和 X 不相关时，式（6）成立。

对于校准能力验证计划，常用 E_n 值评价参加者结果。

g）其他的统计方法，可参见 GB/T 28043 和 IUPAC 国际协议等。

4.4.1.4 需要考虑的其它事项

a）通过参加者结果与指定值之差完全可以确定参加者的能力，对于参加者也是最容易理解的。差值$(x-X)$也称为“实验室偏倚的估计值”。

b）百分相对差不依赖于指定值的大小，参加者也很容易理解。

c）对于高度分散或者偏态的结果、顺序响应量、数量有限的不同响应量，百分位数是有效的。但该方法仍应慎用。

d）根据检测的特性，优先或需要使用变换结果。例如，稀释的结果呈现几何尺度，需做对数变换。

e）如果$\hat{\sigma}$由公议（参加者结果）确定，$\hat{\sigma}$的值应可靠，即，基于足够多次的观测以降低离群值的影响。

f）如果能力统计量（例如E_n值和ζ比分数）需使用参加者报告的测量不确定度的估计值时，只有所有参加者采用一致的方法（比如按照 ISO/IEC 指南 98-3 的原则）评估不确定度，该方法才有意义。

4.4.2 定性结果和半定量结果

4.4.2.1 对于定性结果和半定量结果，如果应用统计方法，必须与结果的特性相适应。对定性数据[也称之为“分类”数据]，可采用直接将参加者结果与指定值进行比较的技术。如果两者相同，则结果是可接受的；如果不相同，可由专家判断参加者结果是否满足预期用途。某些情况下，可审查参加者的结果，并确定该能力验证物品不适于评估，或者指定值不正确。

4.4.2.2 用于定性数据的技术也适用于半定量结果[也称为“顺序”结果]。顺序结果包括很多类型，例如，响应为等级或排序、感官评价，或化学反应强度（如 1+，2+，3+，等）。有时，这些响应结果由数字表示，如，1=差，2=不满意，3=满意，4=良好，5=优秀。

4.4.2.3 对顺序数据，即使结果以数值表示，计算常规的总计统计量是不合适的。因为这些数值并不是基于区间尺度，也就是说，客观意义上，1 和 2 间的差可能与 3 和 4 间的差并不相同，因而不能解释其平均值和标准差的意义。因此，对半定量结果使用诸如 z 比分数的统计量是不合适的。特定的统计量，如秩或顺序统计量，对顺序数据是可以使用的。

4.4.2.4 描述出（或作图表示）所有参加者结果的分布，以及每一类结果的数量或百分比，并给出总计统计量（如众数和极差）是适当的。根据与指定值的接近程度评价结果的可接受性也是适当的，例如，结果落在指定值之上或之下一个数值范围内即为可接受的。某些情况下，利用百分位数评估能力也是合适的，如，可以规定距离众数或指定值最远的 5%的结果是不可接受的。这些规则应根据能力验证计划的目的来确定。

4.4.3 合成的能力比分数

当对一个特定被测量使用了一个以上能力验证物品或有一组相关被测量时，可根据一轮能力验证计划中两个或两个以上的结果评定参加者的能力。这样可以对参加者能力进行全面评定。采用图方法，如尧敦（Youden）图或曼德尔（Mandel's）h统计量图等，也是解释参加者能力的有效工具（参见 GB/T 28043）。

尽量不使用能力比分数的平均值，因为这将掩盖对一个或多个能力验证物品的较差的检测或测量能力，而这正是需要调查的。最常用的合成的能力比分数是可接受结果的数量(或百分比)。

4.5 能力评定

4.5.1 初始能力

4.5.1.1 应根据能力度量方式制定能力评定准则，用于能力评定的方式如下：

a）专家公议，由顾问组或其他有资格的专家直接确定报告结果是否与预期目标相符合；专家达成一致是评估定性测试结果的典型方法。

b）与目标的符合性，根据方法性能指标和参加者的操作水平等预先确定准则。

c）用统计方法确定比分数，其准则应当适用于每个比分数；比分数的常用例子如下：

1）z比分数、z'比分数和ζ比分数（简单起见，示例中仅给出了 z 比分数，对z'比分数和ζ比分数也适用）；

—— $|z|\leq 2$ 表明“满意”，无需采取进一步措施；

—— $2<|z|<3$ 表明“有问题”，产生警戒信号；

—— $|z|\geq 3$ 表明“不满意”，产生措施信号。

2）对E_n值：

—— $|E_n|\leq 1$ 表明“满意”，无需采取进一步措施；

—— $|E_n|>1$ 表明“不满意”，产生措施信号。

4.5.1.2 只要可能，应当使用 GB/T 28043 和 IUPAC 国际协议所描述的图形来显示参加者能力（如直方图，误差条形图，顺序 z 比分数图，尧敦图等）。这些图可用来显示：

a）参加者结果的分布；

b）多个能力验证物品结果间的关系；

c）不同方法所得结果分布的比较。

4.5.1.3 有时，能力验证计划中某些参加者的结果虽为不满意结果，但可能仍在相关标准或规范规定的允差范围之内，鉴于此，在能力验证计划中，对参加者的结果进行评价时，通常不作“合格”与否的结论，而是使用“满意/不满意”或“离群”的概念。

4.5.1.4 当利用测量审核对参加者的结果进行判定时，可利用 E_n 值或参照相关技术标准（包括统计技术方面的标准）进行判定，附录 C 给出了相应的统计方法信息。

4.5.2 长期监测能力

4.5.2.1 能力验证计划可包含长期监测能力的程序。该程序可以使参加者能观测到其能力的变动，是否呈现趋势性的变化或不一致的结果，以及随机变化。

4.5.2.2 图形方法有助于理解数据分析结果，如传统的“休哈特”控制图。数据列表和总计统计量可以提供更详细信息。用来评定能力的能力比分数，如 z 比分数，可用于绘制这些图和表。其它示例和图形工具可参见 GB/T 28043 等。

4.5.2.3 用参加者结果统计得到的标准差作为能力评定标准差时，由于参加者群体的变化及其对比分数的未知影响，长期监测能力时应当谨慎。通常，由于参加者逐渐熟悉能力验证计划或者方法得到改进，实验室间标准差会随时间而减小。即便参加者本身的能力没有变化时，也会导致 z 比分数的明显变大。

附录A

检测能力验证计划常用稳健统计方法

A.1 总则

由能力验证计划参加者的结果确定指定值和能力评定标准差，是检测能力验证计划常用的方法。通常，可以采用经典方法，用格拉布斯（Grubbs）准则等统计方法剔除离群值后计算平均值和标准差，以平均值和标准差作为指定值和能力评定标准差；也可采用稳健统计方法，稳健统计方法不需要用统计方法剔除离群值。例如，使用中位值和标准化四分位距法、GB/T 28043 推荐的算法 A 和算法 S，计算中位值或稳健平均值作为指定值，计算标准化四分位距、稳健标准差或标准差的稳健联合值作为能力评定标准差。本附录描述了由参加者的结果确定指定值和能力评定标准差的常用稳健统计方法。

A.2 算法 A

算法 A 来自 GB/T 6379.5。应用此算法计算得到数据平均值和标准差的稳健值。稳健性是估计算法的特点，而不是其产生的估计值的特点，因此严格来说，称由此算法计算的平均值和标准差是稳健的是不确切的。然而，为避免使用繁琐的术语，“稳健均值”和“稳健标准差”应理解为利用稳健算法计算的总体均值和总体标准差的均值估计。

从一个特定检测中得到的结果总数为 p 。

按递增顺序排列 p 个检测数据，表示为：

$x_1, x_2, \cdots x_i, \cdots, x_p$ 。

这些数据的稳健平均值和稳健标准差记为 x^* 和 s^* 。

计算 x^* 和 s^* 的初始值如下（med 表示中位数）：

$$x^* = \mathrm{med} x_i \qquad (i = 1,2,\cdots, p) \cdots\cdots\cdots\cdots\cdots\cdots (A.1)$$

$$s^* = 1.483 \times \mathrm{med}\left|x_i - x^*\right| \qquad (i = 1,2,\cdots, p) \cdots\cdots\cdots\cdots\cdots (A.2)$$

根据以下步骤更新 x^* 和 s^* 的值。计算：

$$\delta = 1.5 s^* \qquad \cdots\cdots\cdots\cdots\cdots\cdots (A.3)$$

对每个 $x_i (i = 1,2,\cdots, p)$ ，计算

$$x_i^* = \begin{cases} x^* - \delta, & \text{若} x_i < x^* - \delta \\ x^* + \delta, & \text{若} x_i > x^* + \delta \\ x_i, & \text{其他} \end{cases} \qquad \cdots\cdots\cdots\cdots\cdots\cdots (A.4)$$

再由下式计算 x^* 和 s^* 的新的取值：

$$x^* = \sum x_i^* / p \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (A.5)$$

$$s^* = 1.134\sqrt{\sum \left(x_i^* - x^*\right)^2 / (p-1)} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (A.6)$$

其中求和符号对 i 求和。

稳健估计值 x^* 和 s^* 可由迭代计算得出，例如用已修改数据更新 x^* 和 s^*，直至过程收敛。当稳健标准差的第三位有效数字和稳健平均值相对应的数字在连续两次迭代中不再变化时，即可认为过程是收敛的。这是一种可用计算机编程实现的简单方法。

此外，还有一种简化的稳健计算方法可以替代算法 A。按式（A.1）计算稳健平均值，按式（A.2）计算稳健标准差停止，不再进行迭代，以稳健均值和稳健标准差的初始值作为数据平均值和标准差的稳健值。

A.3 算法 S

算法 S 用于计算标准差（或极差），可推出标准差或极差的稳健联合值。算法 S 与算法 A 类似，迭代若干次后最终获得标准差或极差的稳健估计值 W^*。计算的步骤如下：

将 p 个数据以递增顺序排列，表示为：

$w_1, w_2, \cdots w_i, \cdots, w_p$。

（这些数据可以是极差或标准差。）

稳健联合值记为 w^*，每个 w_i 对应的自由度为 ν。（当 w_i 为极差时，$\nu = 1$。当 w_i 为 n 次测试结果的标准差时，$\nu = n-1$。）根据表 A.1，查得算法所需的 ξ 和 η 值。

计算 w^* 的初始值如下（med 表示中位数）：

$$w^* = \mathrm{med} w_i \qquad (i = 1,2,\cdots,p) \qquad \cdots\cdots\cdots\cdots\cdots\cdots (A.7)$$

按以下步骤更新 w^* 的值，计算

$$\psi = \eta \times w^* \qquad \cdots\cdots\cdots\cdots\cdots\cdots (A.8)$$

对于每个 $w_i (i = 1,2,\cdots,p)$，计算

$$w_i^* = \begin{cases} \psi, & \text{若} w_i > \psi \\ w_i, & \text{其他} \end{cases} \qquad \cdots\cdots\cdots\cdots\cdots\cdots (A.9)$$

计算新的 w^*：

$$w^* = \xi\sqrt{\sum \left(w_i^*\right)^2 / p} \qquad \cdots\cdots\cdots\cdots\cdots\cdots (A.10)$$

稳健估计值 w^* 可由迭代算法得到，即不断更新 w^*，直到过程收敛。当稳健估计值的第三位有效数字连续两次迭代后数值不再变化时，即可认为过程是收敛的。这是一种可利用计算机编程实现的简单方法。

表 A1 稳健分析必需的因子：算法 S

自由度ν	限系数η	修正系数ξ
1	1.645	1.097
2	1.517	1.054
3	1.444	1.039
4	1.395	1.032
5	1.359	1.027
6	1.332	1.024
7	1.310	1.021
8	1.292	1.019
9	1.277	1.018
10	1.264	1.017
注：ξ和η值由 GB/T 6379.5-2006 的附录 B 导出。		

A.4 中位值和标准化四分位距法

中位值和标准化四分位距法是一种简单的稳健统计方法。应用此法计算得到数据总体均值和总体标准差的估计值——中位值（med）和标准化四分位距（NIQR）。中位值和标准化四分位距是数据集中和分散的度量，与平均值和标准差相似。

中位值是分布中间位置的一个估计。标准化四分位距等于四分位距（IQR）乘以因子 0.7413。四分位距是高四分位数和低四分位数的差值。对一组由小到大排列的数据，居于中间位置的数据为中位值，有一半的数据高于它，一半的数据低于它；居于下四分之一位置的数据为下四分位数或低四分位数（Q_1），该组数据的四分之一低于 Q_1，四分之三高于 Q_1；居于上四分之一位置的数据为上四分位数或高四分位数（Q_3），该组数据的四分之一高于 Q_3，四分之三低于 Q_3。在大多数情况下 Q_1 和 Q_3 通过数据值之间的内插法获得。

从一个特定检测中得到的结果总数为 p。

按递增顺序排列 p 个检测数据，表示为：

$x_1, x_2, \cdots x_i, \cdots, x_p$。

这些数据的中位值，可按照式 A.11 计算：

$$\mathrm{med}(x)=\begin{cases} x_{(p+1)/2} & p\text{为奇数} \\ \left\{x_{(p/2)}+x_{(p/2+1)}\right\}/2 & p\text{为偶数} \end{cases} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (A.11)$$

可按照式 A.12 和式 A.13 计算 IQR 和 NIQR：

$$IQR = Q_3 - Q_1 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (A.12)$$

$$NIQR = 0.7413 * IQR \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (A.13)$$

注：因子 0.7413 是从“标准”正态分布中导出。

附录B

能力验证计划结果示例

B.1 总则

本附录给出了检测能力验证计划和校准能力验证计划结果示例。其他的更多示例，可参见 GB/T 28043 等。

B.2 检测能力验证计划

能力验证计划可以设计为使用单一样品，有时，为了查找造成结果偏离的误差原因，也可以采用样品对。样品对可以是完全相同的均一样品对，也可以是存在轻微差别的分割水平样品对。均一样品对，其结果预期是相同的。分割水平样品对，其两个样品具有类似水平的被测量，其结果稍有差异。对双样品设计能力验证计划，可按照附录 A 的方法对结果进行统计处理，统计处理是基于结果对的和与差值。

以中位值和标准化四分位距法为例。

假设结果对是从样品对 A 和 B 两个样品中获得的。

首先按下式计算每个参加者结果对的标准化和（用 S 表示）和标准化差（用 D 表示），即：

$$S=(A+B)/\sqrt{2} \qquad D=(A-B)/\sqrt{2} \qquad \text{（保留 } D \text{ 的+或-号）}$$

通过计算每个参加者结果对的标准化和以及标准化差，可以得出所有参加者的 S 和 D 的中位值和标准化四分位距，即 med（S）、NIQR（S）、med（D）、NIQR（D）。

根据所有参加者的 S 和 D 的中位值和 NIQR，可以计算两个 z 比分数，即实验室间 z 比分数（ZB）和实验室内 z 比分数（ZW），即：

$$ZB=\frac{S-\text{med}\,(S)}{\text{NIQR}\,(S)} \qquad \text{和} \qquad ZW=\frac{D-\text{med}\,(D)}{\text{NIQR}\,(D)}$$

ZB 和 ZW 的判定准则同 z 比分数。ZB 主要反映结果的系统误差，ZW 主要反映结果的随机误差。对于样品对，$ZB \geq 3$ 表明该样品对的二个结果太高，$ZB \leq -3$ 表明其结果太低，$|ZW| \geq 3$ 表明其二个结果间的差值太大。

表 B1 为铅精矿中 Cu 的测定结果和统计处理结果。样品 A 和 B 为一对分割水平样品。表 B1 中给出了结果数、中位值、NIQR、稳健变异系数（稳健 CV）、最小值、最大值和极差等统计量。

表 B1 铅精矿中 Cu 的测定结果和统计处理

实验室代码	铅精矿 A w_{Cu}/%	铅精矿 B w_{Cu}/%	S	ZB	D	ZW	方法代码
01	0.927	0.857	1.2615	-3.05§	0.0495	0.35	Cu-1
03	0.952	0.886	1.2997	-0.68	0.0467	-0.12	Cu-1
04	0.977	0.888	1.3188	0.51	0.0629	2.58*	Cu-1
05	0.995	0.921	1.3548	2.74*	0.0523	0.82	Cu-2
06	0.915	0.852	1.2495	-3.79§	0.0445	-0.47	Cu-1
07	0.962	0.900	1.3166	0.37	0.0438	-0.59	Cu-2
08	0.966	0.891	1.3131	0.15	0.0530	0.93	Cu-1
09	0.950	0.889	1.3004	-0.63	0.0431	-0.71	Cu-1
10	0.969	0.901	1.3223	0.73	0.0481	0.11	Cu-1
11	0.949	0.904	1.3103	-0.02	0.0318	-2.58*	Cu-1
12	0.961	0.890	1.3089	-0.11	0.0502	0.47	Cu-1
13	0.940	0.888	1.2926	-1.12	0.0368	-1.76	Cu-1
14	1.02	0.950	1.3930	5.11§	0.0495	0.35	Cu-1
15	0.956	0.898	1.3110	0.02	0.0410	-1.06	Cu-1
17	0.960	0.912	1.3237	0.81	0.0339	-2.23*	Cu-1
18	0.943	0.864	1.2777	-2.04*	0.0559	1.40	Cu-1
结果数	16	16	16	/	16	/	/
中位值	0.958	0.891	1.3106	/	0.0474	/	/
NIQR	0.0143	0.0106	0.01612	/	0.00603	/	/
稳健CV （%）	1.49	1.19	1.23	/	12.72	/	/
最小值	0.915	0.852	1.2495	/	0.0318	/	/
最大值	1.020	0.950	1.3930	/	0.0629	/	/
极差	0.105	0.098	0.1435	/	0.0311	/	/

注 1： 加§号为不满意结果，即 $|z|\geq 3$；加*号为有问题结果，即 $2<|z|<3$。

注 2： 稳健 CV=NIQR/中位值×100%。

图 B1 和 B2 为根据表 B1 制作的 z 比分数序列图。图中按照大小的顺序显示出每个实验室的 z 比分数（ZB 和 ZW），并标有实验室的代码，使每个实验室能够很容易地与其它实验室的结果进行比较。

图 B1 铅精矿中 Cu 的 ZB 柱状图

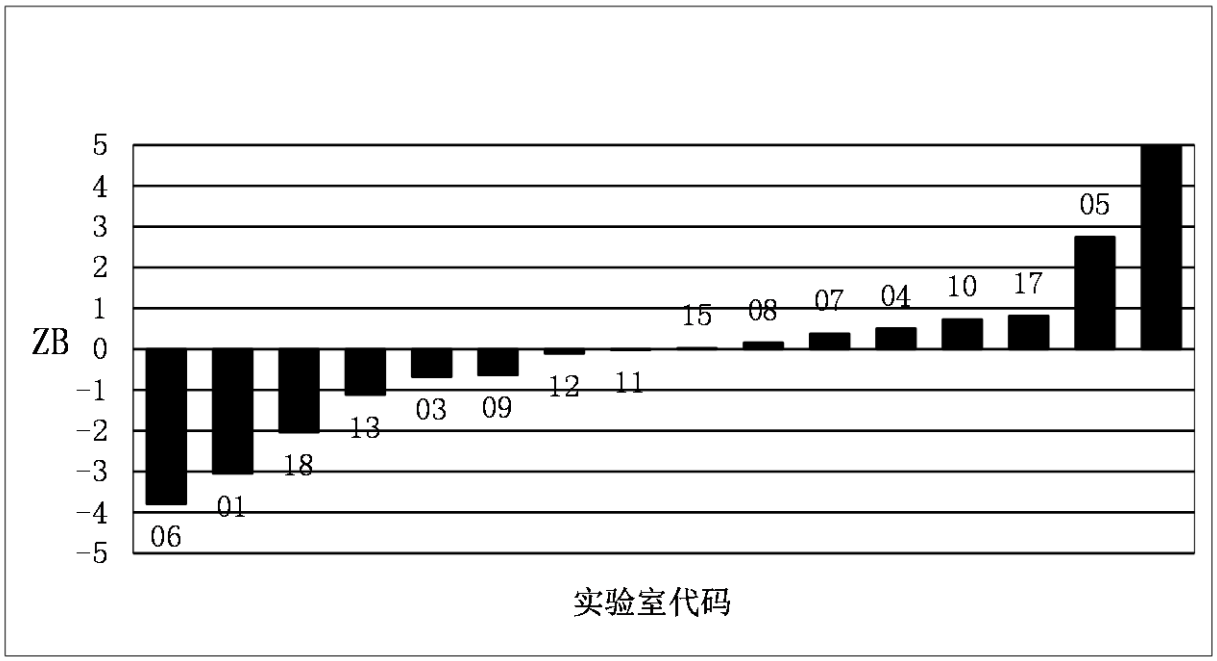


图 B2 铅精矿中 Cu 的 ZW 柱状图

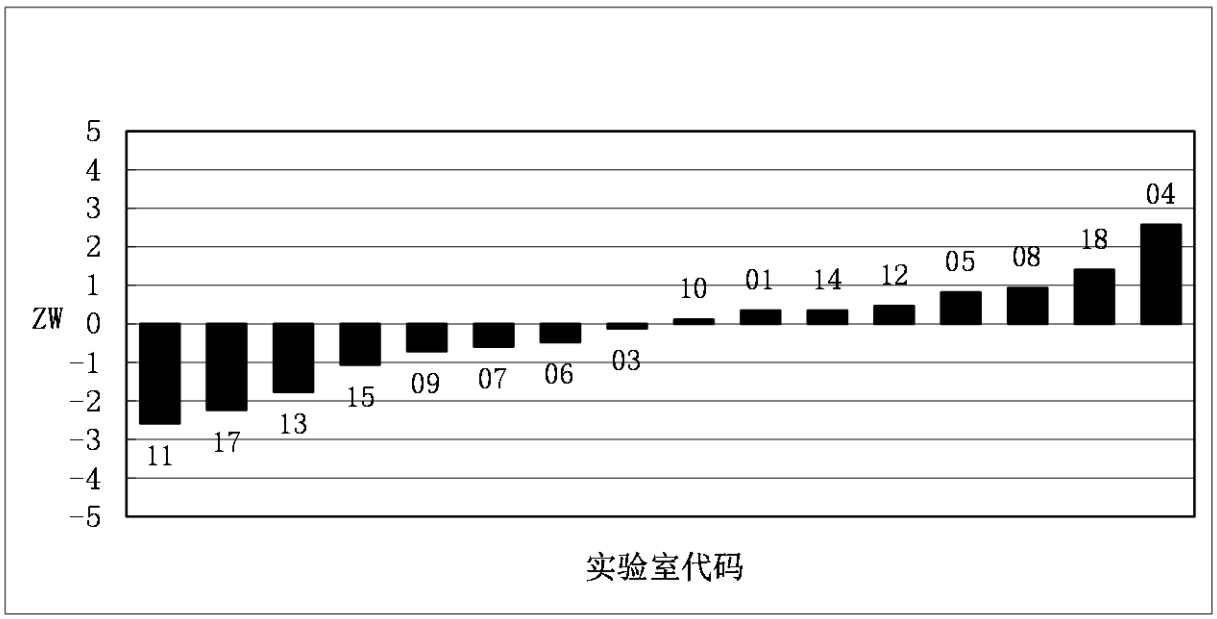
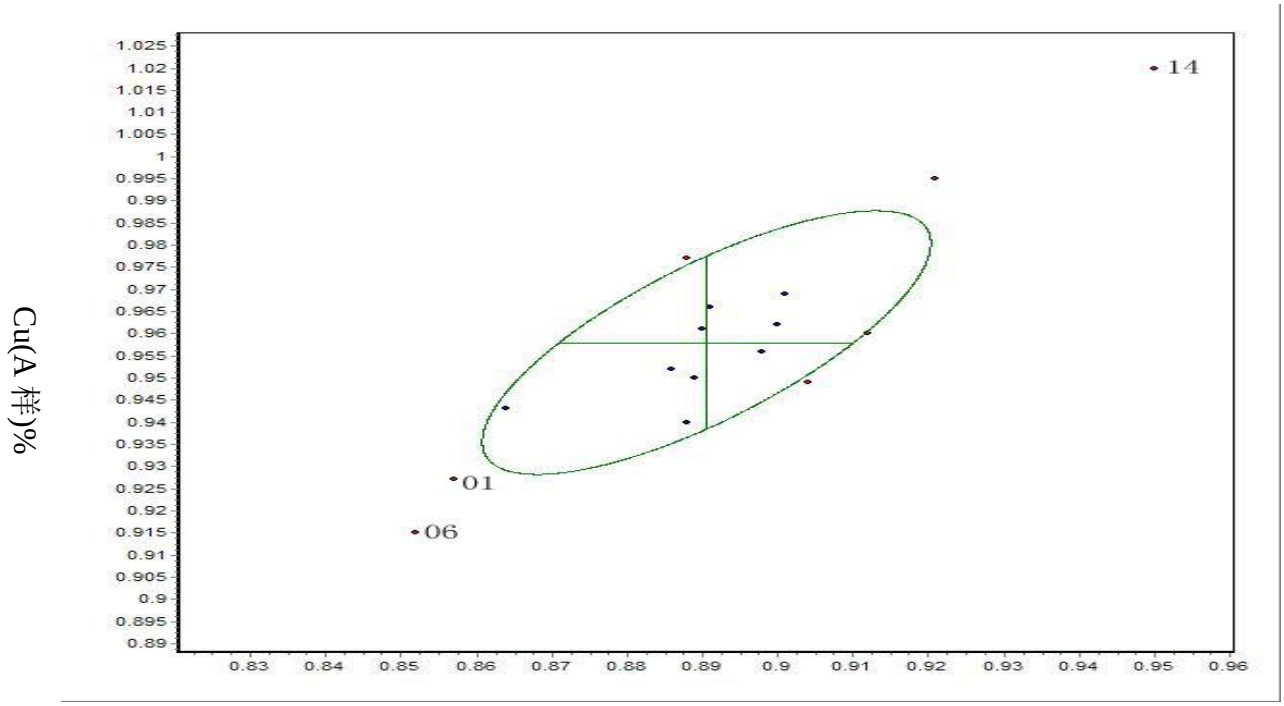


图 B3 为根据表 B1 制作的尧敦（Youden）图。尧敦图是为 2 个样品的结果对而设计的。尧登图能显著地表示出实验室的系统偏差。每个实验室的结果对，用黑点 • 表示。图中的椭圆表示约为 95%概率的置信区域，椭圆的中心为二个样品中位值的交点。

图 B3 铅精矿中 Cu 的分析尧敦（Youden）图



处于椭圆外的所有的点都标有相应的实验室代码。但要注意，这些点并不意味着都是离群。这是因为离群的标准（$|z|\geq 3$）的置信水平约为 99%，而椭圆是约 95%的置信水平。

这意味着，如果数据中没有离群值，期望大约有 5%的结果将在椭圆外。然而因为能力验证的数据通常包含一些离群值，所以在大多数情况下将有超过 5%的点在椭圆外。

图中椭圆外的点，大体相当于那些 z 比分数大于 2 或小于-2 的值。因此，结果在椭圆之外但还不是离群值的实验室应当复查他们的结果。

尧登图的优点在于它们是真实数据的图示。在椭圆外的实验室能够看到它们的结果是怎样不同于其他的实验室。

尧敦图可以说明：

(1) 含有明显系统误差的实验室（即实验室间变异）将在椭圆的右上象限或者在左下象限，即两个样品的结果异常地高或低；

(2) 随机误差（即实验室内变异）明显高于其他参加者的实验室将处于椭圆外的左上或右下象限，即一个样品的结果过高，而另一个则过低。

然而应注意，尧敦图只是用来说明数据，并不用来准确评定实验室的结果（结果的评定仍由 z 比分数确定）。

B.3 校准能力验证计划

在校准能力验证计划中,常使用 E_n 值来评定某一参加者的每一个单独结果。E_n 值并不表明哪个参加者的结果最接近指定值，它只表明其测量结果是否符合参加者声称的不确定度。因此，报告了小的不确定度的参加者,可能和在非常低水平（即较大的不确定度）上工作的参加者具有一个相似的 E_n 值。

在一系列相似的测量中，当考虑 $|E_n|$ 明显大于 1 的结果时，宜评价参加者出具的所有结果，观察是否存在一个系统偏离（例如 E_n 值始终是正值或负值）。

表 B2 为 200 mg 砝码校准能力验证计划结果。为清晰地表示参加实验室的结果，可制作结果图示。以表 B2 中的数据为例：

图示方法一（见图 B4）：每个实验室的实验室偏倚估计值用◆表示，实验室偏倚估计值向上和向下延伸的线段代表实验室偏倚估计值的测量不确定度（$\sqrt{U_{lab}^2+U_{ref}^2}$）。

图示方法二（见图 B5）：每个参加实验室的结果和指定值用◆表示。参加实验室结果和指定值向上和向下延伸的线段代表参加实验室结果和指定值的扩展不确定度。

表 B2　200 mg 砝码校准能力验证计划结果

实验室代码	实验室结果 x /(mg)	实验室结果的扩展不确定度 U_{lab} /(mg)	指定值 X/(mg)	指定值的扩展不确定度 U_{ref} /(mg)	实验室偏倚估计值 $x-X$ /(mg)	E_n 值
001	-0.01	0.01	-0.009	0.004	-0.001	-0.09
002	0.005	0.011	-0.009	0.004	0.014	1.20 §
003	-0.010	0.020	-0.009	0.004	-0.001	-0.05
004	-0.009	0.011	-0.009	0.004	0.000	0.00
005	-0.009	0.017	-0.009	0.004	0.000	0.00
006	-0.011	0.011	-0.009	0.004	-0.002	-0.17
007	-0.004	0.007	-0.009	0.004	0.005	0.62
008	-0.011	0.006	-0.009	0.004	-0.002	-0.28
009	0.00	0.04	-0.009	0.004	0.009	0.22

注 1：不满意结果在表中加“§”标注。

注 2：U_{lab} 和 U_{ref} 的包含因子 k=2。

图 B4　200 mg 砝码校准能力验证计划结果图示（以实验室偏倚估计值作图）

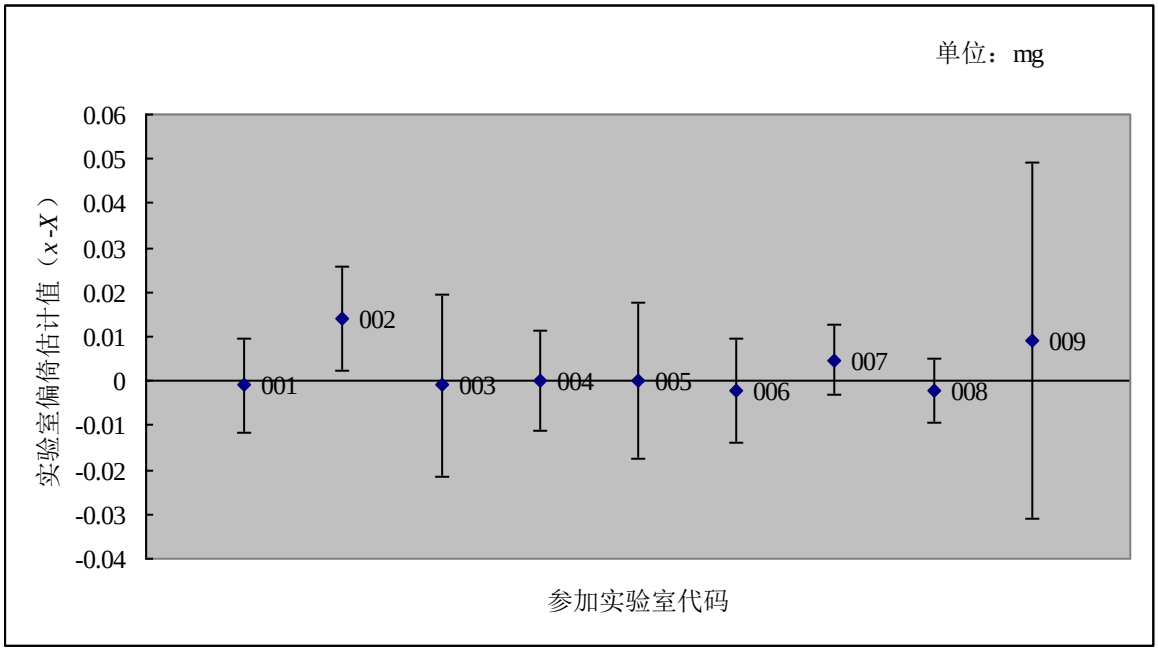
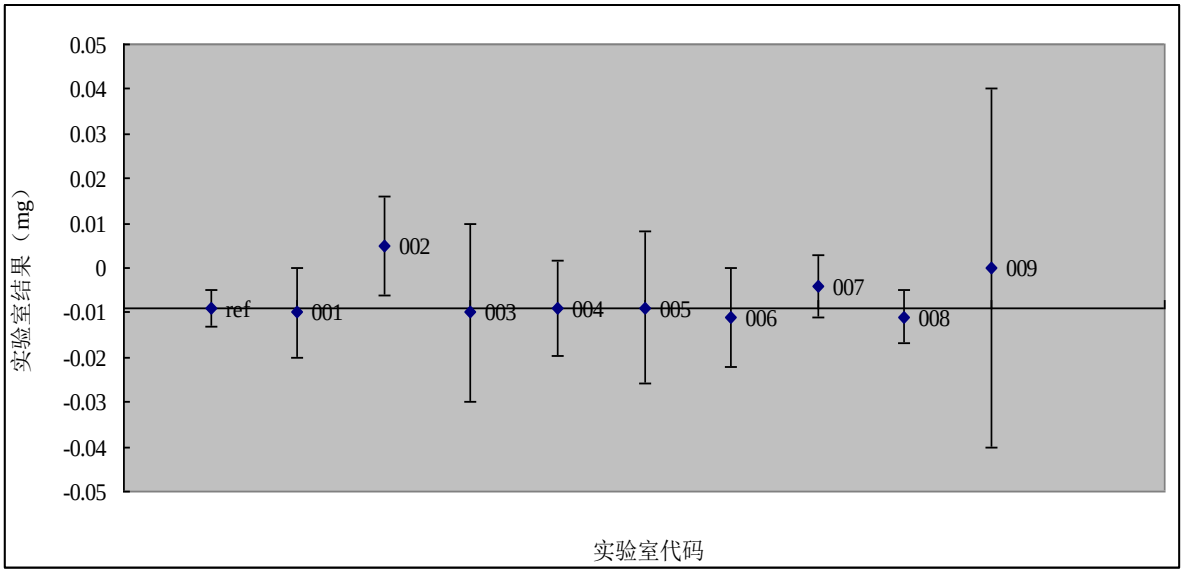


图 B5　200 mg 砝码校准能力验证计划结果图示（以实验室结果作图）



需要注意，图形仅仅是数据的说明，可以宏观地比较所有实验室的结果及其不确定度。它们不代表对一个结果的最终评定（结果的评定由 E_n 值来确定）。

附录C

测量审核结果的评定

C.1 总则

本附录介绍了测量审核结果的几种评定方式。对测量审核结果，可根据参加者、测量方法及测量物品的具体情况，选用合适的方式进行评价。

C.2 测量审核结果的评定方式

C.2.1 按E_n值评定

按 4.4.1.3 中的式（6）计算E_n值。

若$|E_n| \leq 1$，则判定参加者的结果为满意，否则判定为不满意。

利用E_n值评定参加者结果，其前提是参加者必须能正确评定测量不确定度。如果参加者不能正确评定其测量不确定度，则无法使用该方法。

C.2.2 按临界值（*CD*值）评定

当用于测量的标准方法提供有可靠的重复性标准差σ_r和复现性标准差σ_R时，可采用本方法对测量审核结果进行判定。

根据 GB/T 6379.6-2009，按下式计算 *CD* 值：

$$CD = \frac{1}{\sqrt{2}}\sqrt{(2.8\sigma_R)^2 - (2.8\sigma_r)^2\left(\frac{n-1}{n}\right)} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{（C.1）}$$

如果参加者在重复条件下 n 次测量的算术平均值$\bar{y}$与μ_0参考值之差$|\bar{y} - \mu_0|$小于 *CD* 值，则该参加者的测量结果可以接受，结果判定为满意结果，否则判定为不满意结果。

C.2.3 按专业标准方法规定评定

如果相应专业标准规定了测试结果允许差，可按标准规定评定参加者结果。

按下式计算 P_A 值：

$$P_A = \frac{x_{LAB} - x_{REF}}{\delta_E}$$

式中：x_{LAB} -参加者结果；

x_{REF} -被测物品的参考值；

δ_E -标准中规定的允许差。

若$|P_A| \leq 1$，则参加者的结果满意，否则为不满意。