技 術 資 料

試験所間比較における同等性評価についての調査研究 －基幹比較および技能試験の方法論と問題点－

城野克広

（平成21年1月8日受理）

A Survey on the Evaluation of Consistency in Interlaboratory Comparison: The Methodologies and the Problematic Issues in Key Comparison and Proficiency Test

Katsuhiro SHIRONO

1. 緒言

今日，あらゆるモノ・サービス・情報が，非常に活発に世界規模で流通している．今後ますますその量は増大し，その質は向上を求められていくだろう．大量の製品を効率よく検査し，国内への輸入の際にも国外への輸出の際にも，必要な精度で品質を管理するために，測定による値付けと，可能である場合には不確かさによる測定の質の評価が行われるべきである．この値付けや不確かさの報告は，国家標準にトレーサブルに行われることが望まれる．もしも各国間で標準の実現する量が異なれば，その違いは製品の互換性に関わる問題の原因となり，通商取引において技術的な障害となりうる。このために，国際的な取引のある国の間では，国家標準が許容できる範囲で同等である必要がある．とは言え、これは必ずしも新しい話題ではない。世界規模の通商において，計量の統一を図るべくメートル条約が締結されたのは1875年のことである．つまり100年以上も前から国際的同等性についての議論がなされていたわけである．しかしながら，その間すべての国家標準が定義を同じくするSI（The International System of Units，国際単位系）にトレーサブルであることを根拠に，各国間で国家標準について干渉をするということは近年までほとんど行われてこなかった[1]．

今日の国際通商の現状を考えるに，この不干渉は許されない段階に入ってきつつあると言える．食品・医療・運輸というような直観的に安全性・信頼性がきわめて重要な分野を始めとして，あらゆる分野で計測の信頼性が求められている．国際的な貿易が盛んになるつれ，ますますその要求は高くなっていくだろう．ここで問題になるのは，例えば食品中の特定の有害化学物質の濃度が試験項目であったとして，果たして試験結果がどの程度信頼できるものなのかということである．その質が確保されない限り信頼性の要求に本質的には応えていないことになる．各国の標準が独自に維持されている限りでは，試験所の技能以前の問題があり，この問に定量的な結論を与えるのは難しい．

この技術的課題の解決のために計量標準の立場からは，CIPM（International Committee for Weights and Measures，国際度量衡委員会）のもとに運用されている各国NMI（National Metrology Institute，国家計量標準機関）間のMRA（Mutual Recognition Arrangement，国際相互承認協定)[2] であるCIPM MRAと，各国内でのトレーサビリティ体系という仕組みが提案されている[3]．CIPM MRAの3つの大きな柱は，「国際比較」，「品質システム」，「ピアレビュー」である．この協定はCIPMによって動議され，1999年にメートル条約参加の各国のNMIの間で合意された[4] ことを受け，2004年から有効とされるにいたっている．CIPM MRAの主たる目的は先に述べた通り技術的同等性を確認することであり，ひいては署名したNMIが発行する校正証明書の相互承認を提供することである．これが理想的に実現されることで，あらゆる試験結果・校正結果が地理的な条件によらずに受け入れられるという「ワンストップ・テスティング」のための体系が出来上がることになる．もしそれが現実に運用されるなら経済的に大きな効果を生むことは明白である．

CIPM MRAの本文には枠組みが記載されており，2003年の改訂[2] を経て現在に至っている．さらにA-Eの附属書（Appendix）があり，そこには重要なデータや報告

* 計測標準研究部門　物性統計科　応用統計研究室

が追加されていっている．5つの附属書の内容はそれぞれ以下のものである．

A　協定に参加するNMI，そのロゴ及び署名者のリスト
B1　CIPM基幹比較（Key Comparisons）のリスト（報告書が公開されているもの）
B2　RMO（Regional Metrology Organizations，地域計量組織）基幹比較のリスト（報告書が公開されているもの）
B3　RMO補完比較（Supplementary Comparisons）のリスト（報告書が公開されているもの）
C　校正・測定能力（CMC，Calibration and Measurement Capability）の登録量のリスト
D　基幹比較のリスト（試験中，報告書作成中のものも含む）
E　RMOとBIPMの共同委員会（JCRB，Joint Committee of the Regional Metrology Organizations and the BIPM）での取り決め事項

これらのうち附属書A-DはKCDB（Key Comparison Database）[5]において，常に最新のものを確認することができる．附属書Aから，2008年4月の時点でCIPM MRAには70カ国が参加しており，全参加研究所数は200近くにものぼることが確認できる．附属書Bには同時点で600件以上の基幹比較と200件近くの補完比較の結果が公開されている．これらの比較試験の結果を根拠として，附属書Cにその量についてのCMCとともに校正証明書のリストを示すシステムが構築されていて，すでに20000件を超える登録がある．CIPM MRAにより，この校正証明書の相互承認が提供されることにより基幹比較は単なる技能試験以上の性格を帯びている．

今日ではCIPM MRAと同様の考え方は計量の他の分野にも浸透している．試験所認定の分野ではISO/IEC Guide 43-1[6]に取り決める技能試験の結果を，ISO/IEC Guide 43-2[7]に基づいて試験所認定に活用することが行われている[8]．ISO/IEC Guide 43-1に関してはのちに詳述する．法定計量の分野ではOIMLを中心にMAA（Mutual Acceptance Arrangement，型式評価国際相互受入れ取決めの枠組み）が構築されて，2004年にその基本文書が発行されている[9]．これは輸出国の試験機関が発行した証明書を輸入国が信頼し，自国の型式承認の手続きの中でこれを活用しようとするものである．2006年9月末にR60「ロードセル」及びR76「非自動はかり」の2機種を対象にした最初のMAAに基づく相互信頼宣言書（DoMC，Declarations of Mutual Confidence）が署名され，運用されるにいたっている．

試験所間比較の歴史は必ずしも短いものではないが，上に示したように近年特にその重要性を増し，件数も増えていく傾向にある．これまですでにあった問題点や技術的課題はさらに強調され，また規模が大きくなったり，あるいは様々な取り決めがなされたりする中で，新しくいくつかの問題点が浮き上がってきている．国際的動向としてもこのような流れにあって，2007年には第1回となるInternational Proficiency Testing Conferenceがルーマニアで開催され，64件にわたる研究発表がされている[10]．

このために現段階での基幹比較を含む試験所間比較における同等性評価の現状とその問題点についてよく整理をしておくことは，今後の手続き的あるいは統計的問題を解決していく上で非常に有用であると考えられる．このような動機に基づき，本調査研究ではNMIJ（National Metrology Institute of Japan，計量標準総合センター）内での基幹比較・試験所間比較に携わっている職員に，統計的な見地からの問題のみならず，手続き上の問題も含めて幅広に意見を求め，20数名の方から面談あるいはメールの形式で回答をいただいた．また，文献の調査を行うことで現在の試験所間比較における統計的な諸問題の解決への取り組みを調べた．この技術報告においてはこれらの結果を報告する．

この調査研究は5つの章から構成される．第2章においてはCIPMが定める基幹比較の手続きと推奨する統計的方法の概要について簡単に整理する．第3章ではISO/IEC Guide 43に沿った試験所間比較による技能試験の枠組みとその統計的方法について説明する．第4章では今回行った試験所間比較に関するインタビューにおいて見出された問題点について，いくつかの典型的な例を紹介する．第5章においてはこれらの諸問題について，近年行われている統計的な研究の取り組みを紹介する．第6章にまとめを設け，この報告を振り返ることにする．

2．基幹比較の手続きと統計的方法

2.1　基幹比較の手続き

CC（Consultative Committee，諮問委員会）基幹比較の手続きに関してはCIPMからガイドラインが発行されている[11]．このガイドラインは11章からなっており，試験の設計段階から発行段階に至るまでの実際的な手続きが記されている．以下にその概要を示す．番号は簡明さのために付しており，ガイドラインの章立てとは関係

がない.

1. 各CCでのニーズの調査，試験項目，幹事機関の決定.
2. 幹事機関と指定機関によるプロトコルの作成，参加国・回付物・試験形態・スケジュールの決定.
3. 被試験器の回付・試験.
4. 試験が終了した参加者から随時結果（値，不確かさ，備考）を報告.
5. 全ての結果が得られ次第，幹事機関はそれを精査し，異常がみとめられた場合，該当機関に警告，場合によっては撤退を促す.
6. Draft Aの発行.（参加者のみに開示）
7. 参加者からの意見を取りまとめ，KCRV（Key Comparison Reference Value，基幹比較参照値）とDOE（Degree of Equivalence）の報告を含むDraft Bを発行する.（全体に公開）
8. CCからDraft Bが最終報告書として認証されれば，CIPM MRAの附属書BとしてKCDBに登録される.参加者間で合意が得られない，あるいはCCから認証を得られないときにはCIPMに決定が委ねられる.

RMOが主体として実施する比較試験（基幹比較および補完比較）も同様のスキームで行われており，CCの認証を受け，CIPM MRA附属書Bへの登録が行われている.これによりRMOにしか参加していない機関も，CC基幹比較に参加した機関との同等性が認められCMC登録が可能になる.

2.2 基幹比較の統計的方法

基幹比較は前節の手順に沿って実施することが求められている.ただし手順5において何を異常と認めるか，あるいは手順7において，KCRV，DOEをどのように計算するかについてはここには書かれてない.これらに関しては，Cox[12] が提唱しCIPMが非公式に推奨するガイドラインがMetrologia 39号6巻に掲載されており，実際に最近の基幹比較ではよく用いられている.ただし，以下のような場合には、この方法は正当ではないと指摘されている.

- 参加機関の測定の一部，あるいは全部が互いに依存している.
- 移動用標準（traveling standard）が安定でない.
- 一つの移動用標準を参加者間で単純に巡回したという比較の形式をとっていない.
- 参照値が他の方法で与えられることが決定されている.
- 複数の移動用標準が回付され，それぞれが同等に扱われている.
- 波長や振動数のように，参加者が規定された多くの値のそれぞれについて移動用標準を測定する.

また，この文書の中では手順Aと手順Bの2つの方法が提案されており，手順Aの使用についてはさらに以下のことが要求されている.

- 適切な短期安定性ならびに移送中の安定性を実現した移動用標準について，それぞれの機関が測定結果とそれに付随する不確かさを提供していること.
- それぞれの機関の測定は独立におこなわれていること.
- それぞれの機関が計測した測定量の分布として，各々の測定値を平均とし，その付随する標準不確かさを標準偏差とする正規分布が仮定できること

これらの条件が整ったときには，手順Aを適用して一致性の検定を実施することが適切であるとされている.そうでない場合には手順Bを使うことが推奨されている.

(a) 手順A

このガイドラインの中では，一致性の確認はχ^2検定（カイ二乗検定）を用いて行われる.まず重み付き平均yを以下の式を用いて算出する.

$$y = \frac{\sum_{i=1}^{N} x_i / u^2(x_i)}{\sum_{i=1}^{N} 1/u^2(x_i)} \tag{2.1}$$

ここに，x_i，$u(x_i)$は各機関が報告した測定結果と標準不確かさである.Nは参加機関の数である.さらにこの重み付き平均の標準偏差$u(y)$は以下の式から見積もられる.

$$\frac{1}{u^2(y)} = \sum_{i}^{N} \frac{1}{u^2(x_i)} \tag{2.2}$$

観測された$\chi_{\mathrm{obs}}{}^2$値は以下の式から計算される.

$$\chi_{\mathrm{obs}}{}^2 = \sum_{i}^{N} \frac{(x_i - y)^2}{u^2(x_i)} \tag{2.3}$$

自由度ν ＝ N-1のχ^2分布の値が観測された$\chi_{\mathrm{obs}}{}^2$値よりも

大きくなる確率$\Pr\{\chi^2(\nu) > \chi_{obs}^2\}$が0.05を下回ったときには，「すべての報告値は同じ母平均をもち，報告された標準偏差を持った正規分布である」という仮説は棄却される。つまり、一致性は確認されなかったと結論づけられる．逆にこれが0.05を上回ったときには，この仮説が採択されたものと結論づけられる．つまり，一致性が確認され，同時に報告した標準偏差は信頼できるものであると考えられる．($\Pr\{\chi^2(\nu) < \chi_{obs}^2\} < 0.05$の場合には結果の一致が良すぎるため仮説を棄却することになるが，これはほとんど見られない.）また，これと関連する統計量でBirgeレシオが使われることもある．これについては付録を参照されたい.

仮説が棄却された場合の手順は後に示す．検定を通過し一致性が確認された場合，式(2.1)で計算されたyをKCRV x_{ref}として採用する．式(2.2)で与えられる$u(y)$をKCRVの標準不確かさ$u(x_{ref})$とする．これを用いて，参加機関の番号をiとして，その機関のユニラテラルなDOEを値と不確かさのペア$(d_i, U(d_i))$として計算する．

$$\begin{aligned} d_i &= x_i - x_{\mathrm{ref}} \\ U(d_i) &= 2u(d_i) \\ u^2(d_i) &= u^2(x_i) - u^2(x_{\mathrm{ref}}) \end{aligned} \tag{2.4}$$

また，番号iと番号jの機関間のバイラテラルなDOE$(d_{i,j}, U(d_{i,j}))$を以下の式を用いて計算する．

$$\begin{aligned} d_{i,j} &= x_i - x_j \\ U(d_{i,j}) &= 2u(d_{i,j}) \\ u^2(d_{i,j}) &= u^2(x_i) + u^2(x_j) \end{aligned} \tag{2.5}$$

得られた結果とともに，このようにして計算が行われた手順をすべて記録することも求められている．

またもしχ^2検定によって仮説が棄却された場合にはDraft Aをその時点では参加者に配布せず，十分な時間が得られ，経済的にもそうする意味があるのであれば，不一致の原因を調査する．この場合の手順として，以下のものが提案されている(i) 採択されたときと同じようにユニラテラルなDOEを計算して，

$$|d_i| > 2u(d_i) \tag{2.6}$$

となる機関を特定する．(ii) 当該の機関間で正しい結果と不確かさが適切に得られるように議論する．(iii) もしすべての当該の機関が測定結果と不確かさを修正したなら，重み付き平均を計算する手順に戻り，χ^2検定を再び行う．(iv) もし計測結果と不確かさを修正できない機関があり，すべての当該の機関が比較から撤退する準備がある場合，そのデータを取り除いて重み付き平均を計算する手順に戻り，再びχ^2検定を行う．ちなみに，ここでいう「当該の機関」は“laboratory concerned”を筆者が訳したものである．

なお，以下の場合に当てはまるときは手順Bを援用することもありうるとしている．(a) 原因の調査の前にDraft Aが発行されてしまった場合．(b) 外れ値を出している機関が撤退を受け入れない場合．(c) 十分な時間がない，あるいは不一致を解消することによる経済的価値が小さいと考えられる場合.

(b) 手順B

これは各機関が報告した値の分布を仮定して，その分布からKCRVの分布とみなしうるものを再生成する方法である．そのために各参加機関の報告値に分布を仮定する必要がある．十分な情報が得られていない場合，報告値を母平均とし，報告された標準不確かさを標準偏差とする正規分布を仮定すればよい．ガイドラインの中では，相当する情報があるときには正規分布を仮定する必要はなく，ふさわしい分布を用いるべきであるとしている．次にKCRVの再生成の方法を定める必要がある．ガイドラインではメジアンを一つの候補としているが，ロバストであれば他の方法でも構わないとしている．その場合，以下のメジアンという言葉は代わる推定方法に置き換えて読めばよい．次に再生成するKCRVの数Mを定める．ガイドラインでは$M = 10^6$を推奨している．
ここまでの準備の後に，各機関に仮定した分布に従って1つずつの合計N個のランダムな値を発生させる．この組を$x^{(1)}$ベクトルとして，この操作をM回繰り返し、$x^{(r)}$ $(r = 1,\cdots,M)$を記録しておく．つまり，

$$\boldsymbol{x}^{(r)} = \left(x_1^{(r)}, x_2^{(r)}, \ldots, x_N^{(r)}\right)^{\mathrm{T}} \tag{2.7}$$

である．この$x^{(r)}$ $(r = 1,\cdots,M)$について，それぞれメジアンを求めて$m^{(r)}$とする．$x^{(r)}$ $(r = 1,\cdots,M)$についての$N \times M$行列つまり

$$\boldsymbol{Z} = \left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(M)}\right) \tag{2.8}$$

を記録しておく．また，$m^{(r)}$ $(r = 1,\cdots,M)$を以下のqベクトルとして整理する．

$$\boldsymbol{q} = \left(m^{(1)}, m^{(2)}, \ldots, m^{(M)}\right) \tag{2.9}$$

ここでqの成分の平均値をKCRV x_{ref}とする．qの成分の標準偏差を計算し，KCRV x_{ref}の標準不確かさ$u(x_{ref})$とする．またqの成分について95 %包含区間をその区間が最

小になるように決定し，KCRV x_{ref} の拡張不確かさとする．これは最高密度区間と呼ばれる．この決定は計算機を用いて探索的に行えばよい．

参加機関の番号を i として，その機関のユニラテラルなDOEの値と不確かさのペア $(d_i, U(d_i))$ を計算する．

$$d_i = x_i - x_{ref} \tag{2.10}$$

として，$U(d_i)$ については

$$\boldsymbol{r}_i = (\text{row } i \text{ of } \boldsymbol{Z}) - \boldsymbol{q} \tag{2.11}$$

の $\boldsymbol{r}_i$ の成分について，その95 %最高密度区間から定めるとしている．ガイドラインに $U(d_i)$ の最終的な記載方法についての記述はないが，95 %最高密度区間の上限値と下限値を報告するのがもっとも詳細な表記であろう．また，番号 i と番号 j の機関間のバイラテラルなDOE$(d_{i,j}, U(d_{i,j}))$ について，$U(d_{i,j})$ は

$$\boldsymbol{r}_{i,j} = (\text{row } i \text{ of } \boldsymbol{Z}) - (\text{row } j \text{ of } \boldsymbol{Z}) \tag{2.12}$$

の95 %の最高密度区間から定めるとしている．この場合にも，得られた結果とともに，それが得られた方法について記録することが求められている．

3. 試験所間比較による技能試験の手続きと統計的方法

3.1 本報告書で用いる用語について

ISO/IEC Guide 43-1[6] とそれを補足する規格であるISO 13528[13] はそれぞれJIS Q 0043-1[14] とJIS Z 8405[15] としてJIS化されているが，JIS内で用いられている用語の和訳が異なるものがいくつかある．ここでは出版年の新しいJIS Z 8405に基づいた用語を使用することにする．ISO/IEC Guide43-1の紹介の中でもJIS Z 8405に基づいた日本語訳を用いることとする．例えば本報告書中の「コーディネータ」はJIS Q 0043-1中の「調整者」と読みかえられたい．ただしISO 13528の中で“performance statistics”とされているものは，JIS Z 8405中でも「性能統計」,「成績を表す統計量」,また「成績を表す統計指標」と箇所により異なる訳語が現われている．ここでは「成績を表す統計指標」を採用する．ちなみにJIS Q 0043-1では「実績統計量」とされている．

3.2 試験所間比較による技能試験の手続き

試験所間比較による技能試験の手続きについてはISO/IEC Guide 43[6] に規定されている．なおこのGuideはISO/IEC 17043として規格化される準備が進んでおり[16]，ここで紹介した内容についても変更を伴う可能性があることは了解されたい．ISO/IEC Guide 43-1には技能試験スキームの開発および運営に関わるガイドが制定されており，ISO/IEC Guide 43-2には試験所認定機関による技能試験スキームの選定および利用についてガイドが制定されている．ここでは試験所認定にこだわらずに技能試験を洗いなおす観点から，ISO/IEC Guide 43-1に記載された手続きについて紹介する．以下にその流れを簡単に記す．簡明さのために番号を付すが，ISO/IEC Guide 43-1中の章立てなどとは関連しない．

1. 試験所間比較の用途，設計および実行について，企画・調整する際に考慮すべき基本的な原則を規定する．
2. 計画はスキームの開始前に合意され文書化される．統計的設計・試験品目の管理などについても文書化する．（方法・手順については，通常参加者が選択する．）
3. コーディネータが（あるいは外部契約によって）試験品目の準備・調整をし，適切に包装し，参加者に輸送する．順守すべき詳細な指示書を提供する．
4. 試験所から受け取った結果を解析し，付与された値（及びその不確かさ），成績を表す統計指標，実績の評価などについての報告を参加者に返す．
5. 報告書は規定の日程内で速やかに利用可能とする．
6. 実績評価が間違っていると参加者が考える場合に，コーディネータに照会できることが望ましい．また，スキームの進展に活発に寄与するために試験所からのフィードバックを奨励する．

3.3 試験所間比較による技能試験の統計的方法

統計的方法ついてはISO/IEC Guide 43-1にも附属書Aとして記載があるが，その補足として位置づけられたISO 13528に詳しい内容がある．ただし，ISO 13528に示された内容は「技能試験中に得られるデータに対して処置信号または警戒信号を発信するかどうかを判断するための基準となる値またはグラフィックな基準」を示すにとどまり，この信号の発信を持って試験所認定の基準としてはならず，そのために別途基準を設ける必要があることが強調されている．

試験所間比較による技能試験では「付与された値」とその不確かさを定め，「技能評価のための標準偏差」を決定し，「成績を表す統計指標」を算出するという手順が一般的である．「付与された値」と「技能評価のための標準偏差」の決定はコーディネータの責任である．付与された値 X とは基幹比較におけるKCRVにあたる値で

ある．ISO 13528の中ではおおまかに分類すると，「特定品目の定式化（たとえば製造あるいは希釈）によって決定された結果による既知の値を使う方法」，「標準物質を用いた参照値による方法」，「参加試験所あるいは熟練試験所による合意値を使う方法」の3種類が提案されているが，「健全な統計的基礎を備え，スキームの計画を文書によって示している方法がある場合は，その方法を用いてもよい」とされている．

合意値を用いる場合はロバストな平均を使う．ロバストという形容詞は推測するアルゴリズムに付されるべきで，ロバストな平均値というのはいかにも不自然であるが，この規格では便宜のためにその使用を許可している．ロバストな平均x^*とロバストな標準偏差s^*を求める手順は以下のようなものである．(1) p個のデータを，昇順に並び替える．(2) x^*をx_iのメジアンとして，$s^* = 1.483x$（xは$|x_i - x^*|$のメジアン）を計算する．(3) $\delta = 1.5s^*$として，$x_i < x^* - \delta$のとき，$x^*_i = x^* - \delta$，$x_i > x^* + \delta$のとき$x^*_i = x^* + \delta$，それ以外では$x^*_i = x_i$とする．(4) 以下の計算により新しいx^*とs^*を求める．

$$x^* = \sum_i^p x_i^* / p \tag{3.1}$$

$$s^* = 1.134\sqrt{\sum_i^p \left(x_i - x^*\right)^2 / (p-1)} \tag{3.2}$$

更新後にこのx^*，s^*の値が更新前と比べて数字の第3有効数字が変化する場合(3)に戻る．もし変化しなければ，収束したものとみなし，これをロバストな平均値x^*とロバストな標準偏差s^*とする．

付与された値を合意値以外の方法で定めるときにも，得られた値の正当性の評価のためにロバストな平均値x^*は付与された値と比較されることもある．このとき付与された値をXとして，

$$\left|x^* - X\right| > 2 \times \sqrt{\left(1.25 s^*\right)^2 / p + {u_X}^2} \tag{3.3}$$

の場合にはその原因が調査される．ここでu_Xは付与された値の標準不確かさである．

この付与された値とその不確かさに引き続き，技能評価のための標準偏差が定められる．ここでいう標準偏差とは技能試験で得られた値の標準偏差ではない．以下のような方法が提示されている．

1. 例えば法令で定められた許容範囲などの規定値
2. 達成が期待されるレベル
3. 一般的な再現精度のモデルを用いる方法
4. 技能試験以外の精度評価実験から求める方法
5. 技能試験の結果を用いる方法

この標準偏差は次に述べる成績を表す統計指標を定めるために，許容されるばらつきの指標となるものである．技能試験の結果から技能評価のための標準偏差$\hat{\sigma}$を定めたいときには先に述べたロバストな標準偏差s^*を使ってよいとしている．また統計的な基礎を備えた他の方法を使ってもよい．技能試験の場合には結果がある実用的な範囲におさまっていることのみが要求される場合が多いから，必ずしも技能試験の結果をもって標準偏差を算出する必要はない．試験所の報告値が必要以上に揃っていた場合には，適切な技能評価のための標準偏差$\hat{\sigma}$を用いることで，実用上十分許容される技能を持った試験所が処置信号や警戒信号を発信される事態を避けることができる．また，ラウンドごとに技能試験の結果から得られる標準偏差を用いると，成績を表す統計指標は技能評価のための標準偏差の影響を受けるので，複数ラウンドにわたる重要な系統的変化を見逃してしまう可能性がある．この欠点を避けるために，標準偏差をプールして得られるロバストな標準偏差相当の範囲w^*を使用してよいとされる．このロバストな標準偏差相当の範囲w^*を得る方法もISO 13528において標準化されている．

成績を表す統計指標はこれらの値（付与された値とその不確かさと技能評価のための標準偏差）から算出される種々の統計量のことである．まず試験所のかたよりDが以下の式で推定される．

$$D = x - X \tag{3.4}$$

ここで，xは参加試験所によって報告された値である．この値の絶対値や平方を使用することは勧められない．このかたよりがひとつでも$\pm 3\hat{\sigma}$の範囲から外れた場合は処置信号が発信されたと考えるべきである．また試験所のかたよりが$\pm 2\hat{\sigma}$の範囲から外れた場合は警戒信号が発信されたと考えるべきである．ひとつの処置信号が発生したり，また二つのラウンドで継続して警戒信号が発生したりした場合は調査を要する異常が発生した証拠と考えなくてはならない．その他の成績を表す統計指標で重要なものにはz-スコアとE_n数がある．

$$z = \left(x - X\right) / \hat{\sigma} \tag{3.5}$$

$$E_n = \left(x - X\right) \Big/ \sqrt{{U_{\text{lab}}}^2 + {U_{\text{ref}}}^2} \tag{3.6}$$

E_n数におけるU_{lab}とU_{ref}はそれぞれ参加者の結果xと付与された値の拡張不確かさである．z-スコアの解釈は先ほ

どのかたよりに定められた基準と同じで，±2の範囲から外れた場合は警戒信号，±3の範囲から外れた場合は処置信号が発生されたと考えるべきであるということになる．E_n数は通常±1.0を棄却限界として信号を発信する．その他の成績を表す統計指標としてz'-スコア，ζ-スコア，E_zスコアが提案されている．いずれも参加者の結果xに不確かさが付与された場合に利用できる．また特段の注意の下で用いられるのであれば，順位およびパーセンテージ順位などを使ってもよい．

1ラウンドの中で類似した2つの試料を試験する場合，結果を検討するための多くの情報をもたらす図式表現の手法としてYoudenプロットを用いてもよい．これは2つの試料のz-スコアを縦軸，また横軸としてグラフ上にプロットするものである．この棄却域を与える信頼域楕円の使用も標準化されている．また複数回の技能スキームの成績スコアを組み合わせるためには，各ラウンドの不良を明らかにするようにグラフィカルな処理を行うことが望ましい．グラフを使用する方法の一例として，z-スコアを縦軸に，日付を横軸にしたシューハート管理図がある．この場合単一ポイントが処置限界±3の範囲から外れた場合，また3つの連続する点のうちの2つが同一の警戒限界±2の外側に現れる場合には管理はずれ信号を発出することがある．本報告の中では紹介しきれなかったが，グラフィカルな処理の方法は他にもいくつかの方法が標準化されている．

4. 実例の調査

4.1 インタビュー方法

以下4章の内容はすべてNMIJ内の面談によるインタビューによって得られた結果をまとめたものである．まずメールでの簡単な調査を2008年7月中旬に実施し，データ解析において疑問があると思われる点についての調査を行った．データ解析において疑問に思われる点があるとされた方で，スケジュールの都合がつく方に2008年10月までの期間に面談形式でのインタビューを計16回実施し，20程度のケースについて実例を集めた．そのリストを表1に示す．

4.2 実例

CCT-K7（水の三重点）[21), 22)]，APMP.L-K1（ブロックケージの長さ）[25), 26)] の2つの基幹比較と，産業技術連携推進会議知的基盤部会分科会における第50回分析技術共同研究[43)] として実施された精米中の元素分析[43)] を例にとり，ここに詳細に述べる．

表1 本研究におけるインタビューのリスト

国際比較	測温	CCT-K3(ITS-90 of 83K though 933K)[17), 18)], CCT-K5(ITS-90 between the silver point and 1700ºC) [19)], CCT-K7(Water triple point) [20), 21)]
	長さ	CCL-K1(Gauge Blocks) [22), 23)], APMP.L-K1(Gauge Blocks) [24), 25)]
	電気・磁気	CCEM.RF-K5b.CL(Type-N50 S-parameter) (最終報告未提出)[26)], APMP.EM.RF-K8.CL (RF Power) (比較試験実施前)[27)], CCEM-K8 (DC Voltage Ratio) [28)-30)]
	質量関連量	CCM.FF-K1(Water Flow) [31), 32)], CCM.FF-K3(Air Speed) [33), 34)], CCM.FF-K6(Low-Pressure Gas Flow) [35), 36)], CCM.F-K1~ -K4 (いずれも最終報告未提出) [37)]
	物質量	CCQM-P35(Mass fraction of Ethanol in Aquarius Matrix) [38)]
技能試験	化学系	質量分析 [39)-41)], 陽電子寿命 [42)], 産業技術連携推進会議知的基盤部会分析分科会関連 [43)], 日本環境測定分析協会関連 [44)]
	物理系	湿度 [45)]

4.2.1 実例1：CCT-K7（水の三重点）

SI基本単位のひとつである温度は水の三重点で定義されている．この定義が同等に実現されているかを確認するための比較試験がCCTから動議され，2002年から2004年にかけて比較試験が行われた．幹事試験所はBIPMで，他アルファベット順にBNM，CEM，CENAM，CSIR，CSIRO，IMGC，IPQ，KRISS，MSL，NIM，NIST，NMIJ，NMi-VSL，NPL，NRC，PTB，SMU，SPRING，UME，VNIMの全21機関が参加した．試験の方法は“Collapsed-star方式”で行われた．各国の機関はまず移動用の標準を準備し，それと国家標準との比較を行い，その差と不確かさを記録する．その報告値とともに，移動用の標準はBIPMに送付される．BIPMでは全ての移動用標準が二つの参照セルと比較され，値と不確かさを記録されたのち，移動用標準は各国に返送される．また返送されたセルは安定性の確認のために各国の国家標準と比較される．最終的な測定量はBIPMのセルが実現する水の三重点と各国の国家標準が実現する水の三重点の温度差である．

Fig. 1に得られた報告値を値が小さい順に表記している．エラーバーは拡張不確かさ（$k=2$）である．以下，この小節の文章中で機関iといった場合は，このグラフ中でi番目に小さい値を報告した機関を指す．最小の温

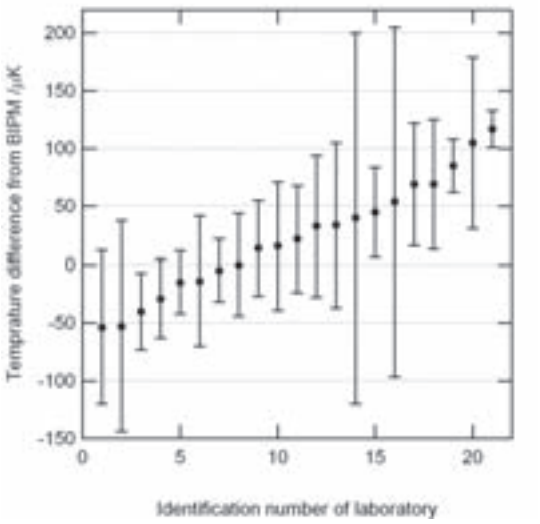

Fig. 1 The reported temperature differences from BIPM derived from Fig. 28 in the final report[20]. Bars show the range of expanded uncertainty (k=2).

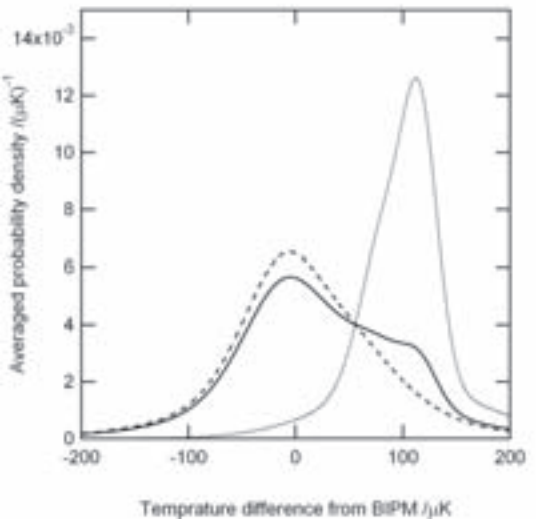

Fig.2 The averaged probability densities from the results shown in Fig. 1 derived from Fig. 29 in the final report[20]. The solid, dashed and dotted lines represent the averaged probability densities of those of all laboratories, 1–18 laboratory and 19–21 laboratory in Fig. 1 respectively.

度差と最大の温度差で,その間の違いは163 μKとなった．これを式(1.3)に代入してχ^2検定を行ったところ，その一致性が確認されなかった．統計的な検討から，大きい値を報告している3つの機関が他のものと有意にずれていることによると判断された．Fig. 2は報告された温度差とその不確かさから，正規分布を仮定して各々の機関のある値における確率密度の平均値を算出したものである．実線が全機関について，破線が値の小さい18機関について，点線が報告した値の大きい機関19–21について平均化したものである．18機関に関してはひとつのモードを持つ正規分布として十分近似できそうなことが確認できる．実際に大きい値を報告した3機関を除いてχ^2検定を行うと残りの18機関の間には一致性をみることができる．また逆に19–21の3機関のみでχ^2検定を行っても一致性を確認することができる．

さらに詳細に検討が行われた結果，大きい値を報告した3機関の報告には共通する特徴があることが分かった．着目するべきは、機関19と機関21はセル中の水の同位体比と不純物濃度に関する定量的情報をもっていたということである．両者ともその値を基にV-SMOW（Viennaの原子力委員会によって準備されている標準海洋水）の同位体比の場合に与えられるだろう結果に補正を行っていた．さらに機関21は不純物濃度についても補正を行っている．また機関20については，複数あるうちの一部の国家標準が高い水の三重点温度を実現しており，V-SMOWに近い同位体比が実現できている可能性が高いことが示唆された．このように他の機関と比べて統計的に有意に大きな値を報告した3機関は海洋水の同位体比に近い水の三重点セルを持っているか，あるいはその値になるように補正をしているという共通した特徴をもっていたと報告された．

機関19と機関21がこのような補正を行う根拠は，CCT-WG8で1990年に出されたITS-90の補足情報[46]内の記述にある．ITS-90の補足情報の2章の中で“an operating triple-point cell contains ice, water, and water vapour, all of high purity and of substantially the isotopic composition of ocean water”なる記述があることから海洋水に補正を行ったものである．この記述に基づいて同位体に関する定量的な評価を行った機関は他にも3つあった．このうち2機関はこれを補正せず，補正をしないことによる不確かさがあるとして補正量を不確かさ成分として報告した．もう1機関は補正をした場合の不確かさのみを報告していて，補正の量は不確かさにもふくまれず，補正も行われなかった．また，その他多くの機関は統計的方法には基づかず，それぞれに同位体比や不純物の影響をBタイプ評価した．幾つかの機関は十分に海洋水に近い同位体比の水を用いるでもなく，しかも補正も不確かさも全く考えなかった．このように様々な処理が行われたために統一的な比較が出来なかったのである．

この場合では同位体比にせよ，不純物にせよ，その補正によって水の三重点が高くなる効果がある．機関19–21の報告した値は他の機関に比べて一方的に大きな

温度差を与える結果となってしまい，しかも不確かさも小さく報告されたから，統計的に一致性が見られないという結果を生むこととなった．

このような経緯を踏まえ，KCRVとしては重みつき平均以外の値を採用することが検討されたが，メジアンも単純平均もそれをKCRVとして採用しても一致性の検定を通過することはなかった．様々な議論の末に，「平均のような古典的な方法によってKCRVを決定する方が好ましいと考える人が多い」というような理由もあり，KCRVとしては単純平均の値が採用された．もちろんこの値はV-SMOWを用いた時の温度差よりも小さい．定量的な補正が可能な機関の値から推定するに70 μK程度の違いがあるのではないかとされた．このような事情から，最終報告書にはKCRVの信頼区間を定めるときには特段の注意が必要である旨が述べられている．この比較試験では特定の機関に同等性を認めずに比較から撤退を促すようなことはなく，全機関に同等性を認めることになった．

この比較試験は結果として，同等性を評価するとか不確かさの算出を共通化するという基幹比較の一次的な目的の枠組みを越え，温度の定義にあったあいまいさを科学的見地から見直す契機となりSIの進歩に貢献することとなった．CCT-K7の結果を踏まえ，SIの温度の定義が変更され同位体比に関する記述がされることになり，2006年に発行されたSIの第8版[47] の中では温度の定義に以下の補足が加えられている．

「補足：この定義は，下記の物質量の比により厳密に定義された同位体組成を持つ水に関するものである：1 モルの^{1}Hあたり0.000 155 76 モルの^{2}H，1 モルの^{16}Oあたり 0.000 379 9 モルの^{17}O，及び1 モルの^{16}Oあたり0.002 005 2 モルの^{18}O.」

同位体比と不純物濃度を考慮した国際比較も現在準備されている．

4.2.2 実例2：APMP.L-K1（ブロックケージの長さ）

長さの標準として重要なブロックケージの校正・測定能力に対する比較試験がAPMPの長さの技術委員会で1998年に動議され，2001年から2002年にかけて実施された．幹事機関はNMIJで，他アルファベット順にKRISS，MSL，NIM，NIMT，NMIA，NPLI，SIRIM，SPRING，VMIの9の試験所が参加した．回付物はスチール製あるいはセラミック製のブロックゲージである．それぞれ長さの異なる10種類ずつの計20本を一組として同時に回付された．試験方法はまず幹事機関で試験を行った後3つの機関で回付され，もう一度幹事機関で測定された．そのあとに残りの6機関を回り，最後に幹事機関でもう一度測定する方式で行われた．この試験方式によって幹事機関のデータは3点得られる．これらから移動用標準のドリフトを見積もることにした．このドリフトは補正されるのではなく不確かさとして考慮された．幹事機関の報告値としては3つの平均ではなく2回目のデータが採用された．また，途中で2つのブロックゲージが深刻な損傷を受けたため，幹事機関の判断でそれ以上の回付は行われなかった．その他，別の2つのブロックゲージについて，最後に測定した機関が表面の傷のために測定できなかったと報告した．

この試験の解析において，まず問題となったのは長期的変動があるかどうかということである．それを簡単に検定するために3度の幹事機関のデータによってχ^2検定を行ってみよう．Fig. 3に例としてセラミック製の呼び寸法80 mmのブロックゲージの測定値を並べた．よく重なっていて，長期変動は見られないようにも見える．実際に相関がないものとして計算すると，$\chi_{\mathrm{obs}}^2 = 3.65$であり，自由度2の$\chi^2$分布の95 %信頼区間の上限値が7.38であるから，長期変動があるとは判断されない．実際にはこれらの長さではドリフトがないと判断される範囲にある可能性が高いが，相関係数が大きいほどにこの検定をパスするのは厳しくなる．特に相関係数を1とすると3つの測定値が完全に一致しているとき以外には，長期変動をしていると判断される．このように，簡単に結論づけることはできないし，95 %信頼限界による線引きに重要な物理的意味があるわけでもない．また他の検定の手法もあって，長期あるいは移送の変動を考慮する基準というものは明確でない．実際には幹事機関の3回の報告値からドリフトに起因する不確かさを見積もった．その不確かさの見積もり方法にも議論があると思うがそれについてはここでは追及しない．

最近、統計的見地から同等性を確認されない値をどのように検出するかということが検討されている．例としてセラミック製の呼び寸法80 mmのブロックゲージに対する呼び寸法からのずれをFig. 4に示す．小さい値が報告されたものから順に示されており，以下この小節の文章中で機関iといった場合はこのグラフ中でi番目に小さい値を報告した機関を指す．この比較試験の報告書では，「(1) E_n数の絶対値を計算する．(2) 最もE_n数の絶対値が大きい機関の値を除き，再度重み付き平均を計算する．それを基に改めてE_n数を計算する．(3) すべての試験所のE_n数の絶対値が1以下になるまで，1試験所ずつ除く操作を繰返す.」という手順で解析が行われた．この逐次E_n数の絶対値が大きい試験所を除いていくとい

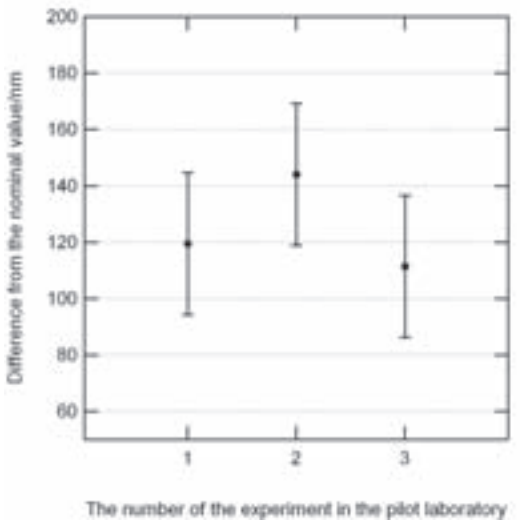

Fig. 3 The measurement results of the length of Ceramic 80 mm (S/N 980260) by the pilot laboratory derived from Fig. 16 in final report[24]. Bars show the range of expanded uncertainty (k=2) without the uncertainty derived from the instability of the traveling standard.

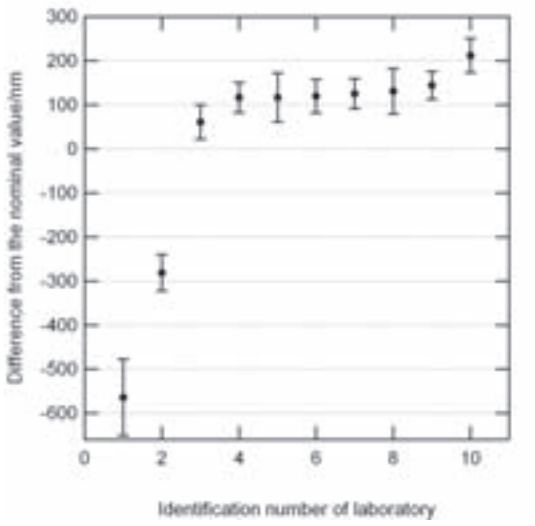

Fig. 4 The results of Ceramic 80 mm (S/N 980260) derived from Fig. 16 in final report[24]. Bars show the range of expanded uncertainty (k=2) including the uncertainty derived from the instability of the traveling standard.

うやり方は2章に示したガイドラインのやり方に沿っているものと考えられる．この比較試験の報告書が2005年にまとめられた後，2007年にCox[48]は必ずしも最大数の機関が一致性を確認できるサブセットがこの逐次的方法で実現されるとは限らないと報告をしている．その報告中で，一致性の取れる最大数のサブセット（LCS: Largest Consistent Subset）が複数あるときにはその中でもっともχ_{obs}^2値を小さくするサブセットを選ぶことを妥当性な選択としている．上の例では実際の手順では，E_n数の絶対値がすべて1以下になるまでに，機関1–3と機関10の4つの機関の値が除かれた．各機関が報告した不確かさに，実際の解析と同様に移動用標準のドリフトに起因する不確かさも加算して，Coxの方法を適用すると機関1，2と機関10の値を除けばχ_{obs}^2値は11.51であり，自由度6のχ^2分布の95 %信頼区間の上側限界は14.45であるから，その値を下回り一致性を見ることができる．これはE_n数の絶対値が大きいものを逐次3つ抜いていったときと変わらない．ただし，4つまで抜く必要はなく，3つの機関を抜けばχ^2検定をパスして重み付き平均を算出する統計的根拠が与えられるということになる．このように外れ値を探す方法は一通りではない．どのような方法が良いかは検証することが困難であり，いずれの方法がよいか優劣をつけるものではないが，特にE_n数の絶対値の大きい方から抜いていく方が手続き上の簡明さがある点は強調されるべきである．

また，このようなRMOの比較試験の結果はCC基幹比較とリンクされることで，RMO基幹比較にしか参加していない機関がCC基幹比較に参加した機関や別のRMO基幹比較のみに参加した機関と同等性の検証をすることができる．しかし、両方の国際比較で同一のブロックゲージを回付することは現実的でない．ブロックゲージの試験においてはドリフトが見られることは先に述べた．また比較試験の最中に移動用標準が深刻な損傷を受けることもある．このリンクのためにDeckerら[49]が現実的な方法を提案している．これは両方の試験に参加した機関について，同様の試験が行われた場合に，番号iの機関がCIPMで$l_{\mathrm{CIPM},i}$，RMOで$l_{\mathrm{RMO},i}$を報告したとき，リンク不変量ι_i

$$\iota_i = l_{CIPM,i} - l_{RMO,i} \tag{4.1}$$

を計算する．さらに，両者の相関係数ρ_iを持って

$$u^2(\iota_i) = u^2\left(l_{CIPM,i}\right) + u^2\left(l_{RMO,i}\right) - 2\rho_i u\left(l_{CIPM,i}\right)u\left(l_{RMO,i}\right) \tag{4.2}$$

を計算して，重み付き平均

$$\bar{\iota} = \frac{\sum_{i=1}^{N} \iota_i / u^2(\iota_i)}{\sum_{i=1}^{N} 1/u^2(\iota_i)} \tag{4.3}$$

とするものである．もちろんχ^2検定を行って一致性を確認すれば，この値の統計的根拠は強くなる．CCL–K1は1998年から1999年にわたって行われた比較試験で

2000年に報告書が発行されており，APMP.L-K1と全く同様の呼び寸法のものが比較されている．3機関がCCL-K1，APMP.L-K1の両方に参加している．上の手順に従ってスチール製のブロックゲージの報告値についてχ^2検定をしてみる．短いブロックゲージではCCとRMOの間での共通のかたよりの成分はそれほど大きくないから，相関をないものと仮定する．この場合，どの長さのブロックゲージにも一致性を確認され，重み付き平均を計算する強い統計的根拠の上にリンクをすることができる．ただし，もし一致性の検定を通過しなかったときには，どのような対処をするのがよいかという潜在的問題は残っている．なお，Deckerらの論文では単純なχ^2検定でなく，後に述べる拡張χ^2検定を用いることを勧めている．

4.2.3 実例3：産業技術連携推進会議知的基盤部会分析分科会：第50回分析技術共同研究（精米中の元素分析）

試験所認定などに使われるような技能試験のデータはここでの公表になじまない．本報告では産業技術連携推進会議知的基盤部会分析分科会の分析技術共同研究を取り上げる．これは一般的な技能試験と異なって，技能の評価というよりも分析技術の向上を目的とした試験であり，以下に挙げるような技術的問題点を見出すというような目的で行われている．従ってここで挙げられる問題点をこの共同研究そのものの問題点として誤解をされぬように注意を喚起しておく．さて，この分析技術共同研究は毎年対象を変えながら，50回目を数える試験所比較による共同研究である．2007年には精米中のCd濃度をはじめとする元素分析について，情報交換や研究に取り組み，国際的に通用する分析技術の確立を図る目的で共同研究が行われた．報告書[43]によれば，日本は火山国であるため，土壌のCd濃度が高く，0.4 ppm以上のCdが含まれる玄米は非常食用に政府が買い取っている．これについて，国際食品規格の作成を行っているコーデックス委員会で上限許容値を0.2 ppmにする案が浮上し，日本が修正案を出し，0.4 ppmに変更されたというような経緯がある．この共同研究のためにNMIJで当時開発中の精米の候補標準物質が用いられ，参加試験所に配布された．参加試験所は元素によって異なり，55から57である．Cd濃度については56試験所から報告があった．技能試験では試験方法は特に定められず，実施者の判断で行われるのが普通である．この共同研究では56件中原子吸光法による報告が13件，ICP発光分析法が39件，ICP質量分析法が3件，蛍光X線分析法－検量線法が1件であった．なお，同一の方法で2回の測定結果を集計の対象とし，その平均値を1件のデータとして取り扱っている．

このデータに対してGrubbsの外れ値検定[50]を適用して4つの報告値を除き，残りの52件に関して統計量を算出することとした．（ただしISO/IEC Guide 43の附属書には，このようにして除外された外れ値も評価の対象ではあるとされているのを根拠に，データ提出数に含め，それらの値についてもz-スコアの算出は行った．）この52個の報告値を昇順に並べたときに25 %，50 %，75 %の順位に相当する値（四分位数という）を$Q1$，$Q2$，$Q3$とする．付与された値としては，$Q2$（つまりメジアン）を採用し，技能評価のための標準偏差としては，

$$s = 0.7143\left(Q3 - Q1\right) \tag{4.4}$$

を用いた．この四分位数を用いた標準偏差の推定は分布が正規分布で十分な参加試験所の数があるときに正当な方法であり，試験所間比較ではよく使用される．これを用いてz-スコアを算出した．上に示したようにz-スコアが±2の範囲から外れた場合は警戒信号が発せられると考える．逆に$|z|$が2以下であれば満足な結果であるとする．この比較試験においては$|z|$が2以下となった試験所は41試験所であった．6試験所が$z = -2$よりも小さく，9試験所が$z = 2$よりも大きかった．その付与された値は，0.190 mg/kgで，ばらつきの標準値は，0.015 mg/kgであった．NMIJで均質性試験のみから得られた分析値は，0.195 mg/kgで均質性試験のみの標準偏差は0.003 mg/kgであるが，不確かさはこの時点では十分に検討されていない．ちなみに先に紹介したロバストな平均と標準偏差を計算すると$x^* = 0.190$ mg/kgと$s^* = 0.017$ mg/kgで，ほとんど同じ結果が得られる．

付与される値にせよ，技能評価のための標準偏差にせよ，正規分布を根拠としている．実際に分布が正規分布で，用いた平均と標準偏差が適切なものならば，$|z| < 2$は95 %信頼限界に値が収まっていることを意味している．しかし52試験所の5 %は2, 3試験所にもかかわらず，実際にはz-スコアの絶対値が2以上とされた報告値は（外れ値とされた4つを除いて）11にも上る．Fig. 5は縦軸をその値の順位を（試験所数 +1）で割ったうえ，累積正規確率分布の逆関数をとったもので，横軸は報告値である．これは正規分布なら直線になり，黒線が付与された値0.190 mg/kgとばらつきの標準値0.015 mg/kgをそれぞれ平均と標準偏差とする正規分布の場合に予想される直線である．実際の試験の結果得られた点の分布は黒線と比べるとややS字の傾向を示す．S字は正規分布より

も裾の広い分布に特徴的な形である．この結果からはメジアンを平均とし四分位数から求まる標準偏差を母標準偏差とする正規分布ではなさそうであり，もっと裾の広い分布と考えるのが妥当であろう．実際にGrubbsの検定で除いたのちの52の報告値の標準偏差を計算すると0.015 mg/kgよりもかなり大きい．この共同研究の報告書にも「*z*-スコア法はその原理が正規分布を前提としたものであり，実際の系に適用した場合は技能試験を測定する尺度としては参考となる数値でしかない」旨が述べられている．

ここでは原子吸光法とICP発光分析法で実験したグループが多かったが，グラフに示していないGrubbs検定でのぞかれた4つのうち3つの値は原子吸光法であり，この方法により得られた値はほかの分析法により得られた値より大きかった．（残りの一つはICP発光分析法であり，この方法ではCdは検出されなかった．）これらのことを合わせて考えると，各試験所が用いた標準や試験方法に起因するかたよりが生じていることも考えられる．あえて両者を分けて上記と全く同様の手順で*z*-スコアを評価されたと仮定すると，グループ分けする以前には*z*-スコアが2以上と判定された一部の試験所の*z*-スコアが2以下になる．ただし，両者の母平均に有意な差があるかという検定を行っても，5 %有意水準では有意な差があるとは判定されない．

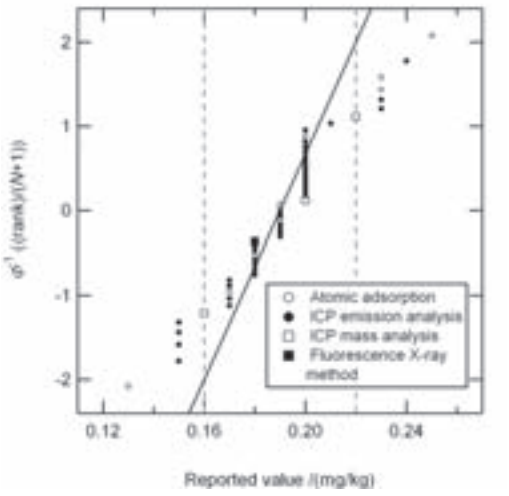

Fig. 5 The normal probability plots for the results of the interlaboratory comparison for the analysis of Cd concentration in rice. Black line shows the plot for the normal distribution with mean of 1.90 mg/kg and standard deviation of 0.015 mg/kg. Dashed lines show $|z|=2$. White and black circles are the results with atomic adsorption and ICP emission analysis. White and black squares are the results with ICP mass analysis and fluorescence X-ray method. 52 results except for 4 outliers are plotted.

このような状況が生じることもあり，比較試験を行う場合にはその目的を明確にしておくことが重要である．ここで紹介したように分析技術の向上を一義的な目的としているときなら，結果をもとに異なる実験方法に対する物理的考察が加えられることもあり得るだろう．一方で認定や技能の評価が目的である場合には，上のような枠組みではなく，認証参照値を付与された値として試験を行うことが好ましいかも知れない．付与された値を熟練試験所間の比較で定め，標準物質として特性値を付与したいとか，値を付与するのはNMIや認証試験所であるが，熟練試験所間の比較によって測定プロトコルの妥当性を確認したい[42]というような目的の場合には，熟練試験所の選定には特に注意すべきである．技能の比較を目的にする場合でも，認証値の存在する対象は試験の性格上好ましくないと考える意見もあり，その際，熟練試験所による合意値とその不確かさを予備的に求めることが行われることがある．熟練試験所の選定をどのように行うかは今後重要な課題である．

4.3 問題点の整理

他にも多くの基幹比較，技能試験の例を見た．問題となる点は共通的なものが多い．以下のように整理できるのではないかと考えた．

(a) 不確かさや補正量の算出

CCT-K7の例でみたように，重要な不確かさや補正の要因を見積もるのが難しかったり，見逃したりすることは多々ある．不確かさが過小に評価されると同等性が確認されなくなるおそれがある．化学系の試験では繰返しのばらつきが大きく，そこに注意が集中してしまうために，技能試験では不確かさが過小評価されることが問題であるようである．このような状況を改善するために，日本分析化学会が主催して，「トレーサビリティと不確かさ」理解のための分析技能試験[51]なども行われている．また別の問題として，液体流量の基幹比較CCM.FF-K1[31), 32)]ではAタイプの評価が通常とは異なる方法で幹事により指定されたことがあげられる．本質的な問題ではないかも知れないがここに分類しておく．

(b) 同等性の評価・外れ値の検出

同等性の評価は基幹比較を含む試験所間比較の主たる目的である．その方法は具体的に提案されているが，KCRVを算出することを含め，結果整理・内部報告・一般公開する手順において，同等性が認められないものを合理的に指定し，合意形成するためのガイドラインはま

だ作られていない．DC Voltage Ratioの基幹比較CCEM-K8[30]においては，かなり複雑な議論の末に，いくつかの試験所の報告値はKCRVを求める計算には使わないことを決めた．（このようにKCRVの計算に一部の試験所の値が使われないことはよくある．しかし，それによって，それらの試験所の同等性まで否定したことにならないと解釈するのが妥当だろう．附属書Bに登録されている以上，それらの試験所の測定結果が同等ではなかったとの特記を伴っていない限り，同等性は確認されたというのがCIPM MRAにおける見解であると筆者は認識している．）同等性の評価の方法には個々の分野や試験の性質に応じてさまざまな問題があるので，必ずしも一律のガイドラインが必要ではないが，比較試験の種類に応じて簡明な決定方法の提案が期待される．

(c) 複数被試験器や基幹比較間のリンク

CIPM MRAには70もの国と200以上の試験所が参加している．CC基幹比較は本来他の国に標準を依存せず，SIの定義に則って標準を実現できている試験所が参加すべきものであるので，CC基幹比較だけでは全ての国の間で同等性の確認をすることはできない．従って，RMO基幹比較の役割は非常に重要である．RMOの結果はCCの結果とリンクされたのち，どちらかにしか参加していない機関の間での同等性の評価が可能となる．その実現のためにはリンクをどのようにするかということがひとつの大きな問題である．また同一の比較試験内でも，参加機関の間でグループ分けが行われ，それぞれのグループで別々の標準を用いて試験を行う場合にも同様の問題がある．

(d) 被試験器のドリフト・長期不安性・移送の不安定性

力学量関連の比較試験では仲介器の長期的あるいは移送の不安定性が非常に大きな問題であり，その変動に起因する不確かさあるいは補正を考えないと統計的な一致性は通常確認されない[37]．このような場合にCIPMが示すガイドラインを適用することは難しい．不確かさに含めるのと補正を行うのとでは，統計的位置づけは変わるので，その物理的性状もよく考慮して対処する必要がある．その解決方法を見出すことは今後の統計的方法の研究における重要課題である．

(e) KCRV、付与される値

KCRVや付与される値は一致性の検定，成績を表す統計指標の算出に使われるので非常に重要である．報告値とその不確かさが信頼できるときにはKCRVを重み付き平均とする方法が強い正当性を持っている．報告値や不確かさが信頼できないときに，理論的根拠を持つ方法はかなり特殊な場合にしか見当たらない．基幹比較ではそのような状況に対応するために手順Bが用意されているが，現実のデータ解析においては手順Bのように複雑なものはあまり好まれていないようである．

CCT-K7の例で見たように，重み付き平均を使いたくない状況では単純平均を使いたいという要求があるようであるが，これは理論的な裏付けが弱い．（Kackerら[52]の研究のように，それに理論的裏付けを作ろうとする研究もある．）なおKCRVが重み付き平均でない値y^*のときには一致性の検定は，

$$\chi_{\mathrm{obs}}^{\ 2} = \sum_{i}^{N} \frac{\left(x_i - y^*\right)^2}{u^2\left(x_i\right) + u^2\left(y^*\right) - 2\,\mathrm{cov}\left(x_i, y^*\right)} \tag{4.5}$$

として，自由度N（参加機関数）のχ^2分布と比較することができる[53]．この場合，自由度の数が式(2.3)と，変わっていることに注意されたい．

技能試験においてロバストな平均とロバストな標準偏差はよく使われるが，必ずしもいつも理論的に信頼できる背景を備えているわけではない．その他，熟練試験所による合意値を定めたいときにその選定に問題があることはすでに述べた通りである．

実際の運営ではここに分類した(a)-(e)のような諸問題が複合的に生じるものである．このように予期しうる形でも予期しない形でも様々な問題が起こりうるから，比較試験の計画は慎重に立てなくてはならない．また試験所間比較による技能試験を試験所認定に使用する場合にはここで示したことと異なる種類の問題が生じる可能性がある．

5. 文献調査

上記のような問題を統計的な見地から解決すべくいくつかの方法が考えられている．分類の仕方は見やすさのために4.3節の分類とは違えている．それらの対応については付録1を参照されたい．

5.1 相関がある場合

相関がある場合には基幹比較のガイドラインをそのまま使用することができない．この場合には，各機関の算出する母分散のみならず，二つの機関間の母共分散が必要であり，母分散を対角成分，母共分散をそれ以外の成分とする分散共分散行列Vが予め求められている必要が

ある．もし，共通の標準を用いているなどで，共通の不確かさでかたよりの成分がu'であり，それ以外の全ての標準不確かさを合成したものがu_i'，u_j'である場合，相関係数$\rho_{i,j}$の推定値$r_{i,j}$が

$$r_{i,j}=\frac{u'^2}{\sqrt{\left(u'^2+u_i'^2\right)\left(u'^2+u_j'^2\right)}} \tag{5.1}$$

と推定されることを用いれば，解決される問題は多いだろう．

同じ不確かさ要因であっても，たとえば標準の長期的な変動のように，ばらつきの成分であれば，u'に含めてはいけない．分散共分散行列の各成分$V_{i,j}$は，各機関の出す報告値の標準不確かさをu_iとすれば，$V_{i,i}=u_i^2$，$V_{i,j}=r_{i,j}u_iu_j$ $(i\neq j)$で与えられる．これを踏まえ，Vの逆行列V^{-1}の(i,j)成分を$V_{i,j}^{-1}$とするならば，重み付き平均は，

$$\begin{aligned} y&=\left(\boldsymbol{1}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{1}\right)^{-1}\boldsymbol{1}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x}\\ &=\frac{\sum_i^N\left(x_i\sum_j^N V_{i,j}^{-1}\right)}{\sum_i^N\sum_j^N V_{i,j}^{-1}} \end{aligned} \tag{5.2}$$

で与えられ，

$$\begin{aligned} \chi_{\mathrm{obs}}^{\ 2}&=\left(\boldsymbol{x}-y\boldsymbol{1}\right)^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{x}-y\boldsymbol{1}\right)\\ &=\sum_{i=1}^N\sum_{j=1}^N\left(x_i-y\right)V_{i,j}^{\ -1}\left(x_j-y\right) \end{aligned} \tag{5.3}$$

を計算すれば，手順Aと数学的に等価なχ^2検定を実施することができる．DOEについての定めは当然ないのであるが，χ^2検定をパスし，$x_{\mathrm{ref}}=y$とされるのであれば，ユニラテラルなDOE$(d_i, U(d_i))$は

$$\begin{aligned} d_i&=x_i-x_{\mathrm{ref}}\\ U\left(d_i\right)&=2u\left(d_i\right)\\ u^2\left(d_i\right)&=u^2\left(x_i\right)-u^2\left(x_{\mathrm{ref}}\right) \end{aligned} \tag{5.4}$$

とするのが自然である．ここで

$$\frac{1}{u^2\left(x_{ref}\right)}=\left(\boldsymbol{1}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{1}\right)=\sum_i^N\sum_j^N V_{i,j}^{\ -1} \tag{5.5}$$

である．バイラテラルなDOE$(d_{i,j}, U(d_{i,j}))$は，

$$\begin{aligned} d_{i,j}&=x_i-x_j\\ U\left(d_{i,j}\right)&=2u\left(d_{i,j}\right)\\ u^2\left(d_{i,j}\right)&=u^2\left(x_i\right)+u^2\left(x_j\right)-2r_{i,j}u\left(x_i\right)u\left(x_j\right) \end{aligned} \tag{5.6}$$

とするのが自然であろう．

Willink[54]は相関のない場合も含めDOEの解釈についての詳細を検討しており，有効自由度を考慮した場合のDOEなどとの比較を行っている．またCoxはこのように相関を考慮しないことによる弊害について文献55で紹介している．一方, Douglasら[53]はこのχ^2検定をさけ，Pair-difference χ^2検定という異なる方法を提案している．

5.2 非正規分布の取扱い

現在の基幹比較のガイドラインの手順Aを用いる場合には通常は正規分布が仮定される．しかし，有効自由度を伴ったt分布として報告されることも多い．場合によってはt分布すら不適切であろう．t分布で有効自由度が小さいときや，正規分布やt分布が不適切な場合には一致性の検定をχ^2検定により実施するのは統計的な根拠が薄弱である．これに代わる検定として，拡張χ^2検定がSteeleのグループから提案されている[56]．この検定法は5.1節で述べたPair-difference χ^2検定を元にしている．また同じく一致性の検定に対する改良をIyerら[57]も提案している．これもアプリオリに正規分布を仮定せず自由度に関する情報を利用するものである．Zhang[58]は繰り返し回数に関する情報を利用することで，重み付き平均の不確かさがより大きくなることを指摘している．またWillink[59]はバジェット表から得られる分布の形を想定し，KCRVの分布の形としてはすべての参加機関の分布を掛け合わせるものがよいとしている．通常自由度を考慮すると正規分布を仮定したときより分布が広がるから, 一致性の検定を通過しやすくなるであろう．ゆえに，これらの方法は5.5節で述べる一致性の検定で一致性の仮説が棄却された場合の対処としても有効かも知れない．

5.3 複数ループあるいはCC基幹比較-RMO基幹比較間のリンクついての研究

4.2.2節で紹介したDeckerら[49]の方法が筆者としては，信頼できる方法ではないかと考えている．しかしながら，先にも述べたとおり，一致性が確認されなかったときに，どのような対処があるかは未だ問題である．また別の問題点として，CC基幹比較で行われた試験とRMO基幹比較で行われた試験の内容が同一とは言い難いときに，似たような試験の結果をどのように援用するかということがある。一定の提案はDeckerらの論文中でされているものの、疑問を残す．またRMOにしか参加していない機関には，リンク不変量の不確かさが，機関の技術的レベルとは関係なく加味されるが，CCにしか参加

していない機関には加味されないという不公平があり，同一の比較試験内で複数の同等なループがあるとき[19]のリンク方法としては不十分である．ほかに提案されている方法のうち，Sutton[60]らの方法はCC基幹比較のKCRVを真値と認めることで，解析しているが同じ問題がある．また，いずれの場合も一致性の検定は可能であるが，一致性が見られないときの対処の仕方がないことが問題である．これについては5.5節の方法が適用できるのかも知れない．その他の方法については文献60中の参考文献を参照されたい．

5.4 ドリフト，長期安定性の問題

線形のドリフトについては，いくつかの検討がなされている[61]が，長期や移送の不安定性がある場合の解析方法の提案は少ない．Elsterら[62]は，スター形式で試験が行われたときに解析する方法を提案している．しかしながら，連続する試験は比較しやすいが，遠くはなれた試験は比較しにくいという問題点は克服されておらず，まだ改善の余地はある．

5.5 一致性が見られなかった場合の統計的方法

先の問題点でも見られたように，最大数の機関が一致性を確認されるサブセットを探索的に見出すという方法がCox[48]から提案されている．整理し直すと，χ^2検定によって仮説が棄却された場合に，$|d_i| > 2u(d_i)$なるすべての機関が撤退しても続くχ^2検定を通過するとは限らない．逆に最初の段階で$|d_i| > 2u(d_i)$なるすべての機関が撤退する必要がないときもある．このため，$|d_i/2u(d_i)|$が大なる機関から逐次1つずつの機関が撤退したものと考えて，全体としてχ^2検定を通過するサブセットを選ぶことを行うことが考えられる．一方で，その手順では必ずしも最大数の機関の間で一致性の検定を通過する組み合わせが実現されるとは限らないという指摘がある．そこで，最大数の試験所が参加し，一致性の検定を通過するサブセットを探索的に決定するのが妥当であると上記の論文では提案している．

また一致性がとれていない場合には，不確かさの見積もりが不十分であったり，補正できていないバイアスがあったりすることが考えられる．これらの不確かさやバイアスをある仮定のもとに推定する統計的方法が提案されている．基本的に比較試験の結果を用いて，バイアスや未知の不確かさを見積もるという手順で実施される．古くに提案され，使用の実績もあるのはPauleとMandelの方法[63]である．これは共通の不確かさσ_{com}を与える方法である．観測される$\chi_{\mathrm{obs}}{}^2$値を以下の式で与えられるとする．

$$\chi_{\mathrm{obs}}{}^2 = \sum_i^N \frac{(x_i - y)^2}{u^2(x_i) + \sigma_{\mathrm{com}}{}^2} \tag{5.7}$$

$$y = \frac{\sum_i^N x_i / \left[u^2(x_i) + \sigma_{\mathrm{com}}{}^2\right]}{\sum_i^N 1 / \left[u^2(x_i) + \sigma_{\mathrm{com}}{}^2\right]} \tag{5.8}$$

ここで自由度νのχ^2分布の期待値がνで与えられることから，

$$\nu = N - 1 = \sum_i^N \frac{(x_i - y)^2}{u^2(x_i) + \sigma_{\mathrm{com}}{}^2} \tag{5.9}$$

を解くことで，共通の不確かさσ_{com}を求める．yもσ_{com}の影響を受けるためこの式は数値計算で解かれるのが普通である．欠点としては，この方法を用いるとχ^2分布の値が$\nu = N-1$という限られた値しか与えないことがある．実際には$\chi_{\mathrm{obs}}{}^2$値はχ^2分布に従う．その分布を考慮した場合，重み付き平均の値自体はそれほど大きく変わらないが，重み付き平均の不確かさはかなり変動することがあるということを指摘しておく．

Willinkは[64]最尤法を用いて，共通の標準偏差を求める別の方法を提案しているが，これも最尤推定値のみを採用することには同様の問題がある．また，WeiseとWögerは[65]一致性の問題を解消するために，4つの方法を提案している．彼らは，事後的に個々の報告値に追加的な不確かさとバイアスを，KCRVとの乖離の度合いから推定する方法を推奨している．個別に不確かさやバイアスを推定できることがこの方法の利点である．この場合にはKCRVの値は操作の前後で変わらない．ただし，この方法を用いてもχ^2値が限られた点しか考慮しないという問題点は残る．

5.6 ベイズ統計を用いる方法

近年統計的手法としてベイズ統計を用いる機会が増えている．Kackerら[66), 67)]の一連の論文によって，ガイドラインに定められた手順Aの統計手法がベイズ統計の見地から整理され直した．与えられる形式はガイドラインで示されたものと変わらないが，比較試験を行う前に確率分布に関する事前情報があるという場合などには非常に有効であると期待される．

6. まとめ，今後の課題

一口に試験所間比較といっても，様々な様式が存在し

ている．その性質や試験を行う意義には全体に共通したものとそうでない部分がある．例えば基幹比較では全員の一致が見られなければならないが，通常の技能試験では，参照試験所に対して1対1の同等性が取れれば十分な場合もある．これは統計的見地からはかなり大きく異なる設計が必要である．またどのような様式の比較試験でも，それぞれの量に依存する部分は大きい．細部に関しては統一的な枠組みでは議論されていないし，そうである必要もないと感じた．統一的なプロトコルの整備とともに，むしろ試験後の解析手順を含めて，比較試験ごとに計画を慎重に設計することは非常に重要であろう．相互承認の枠組みが有効となったのが，2004年であるから本格的な取り組みは最近の数年しか行われていない．その意味で様々な問題点があげられるのはある意味では当然であり，むしろその短い間に方法論は非常に洗練されてきているように感じる．

また技能試験では不確かさの算出が未だ信頼できるレベルにないという点で基幹比較とは状況が異なっている．このように異なる枠組みの中では必要とされる統計的方法も異なるかも知れない．試験結果についての新しい方法を提案すると同時に，不確かさ算出の技術的普及を図ることも技能試験のための試験所間比較の信頼性を高めるために重要であろう．

5章では個別的な問題に有用と思われる統計的手法の整理を行った．長期不安定性・移送の不安定性を除けば，それぞれの問題に対して多くのオプションが提案されている．一方ではそれらをどのような場面で使用するのかという問題は残されている．今後も機会あるごとにこの問題に対する統計的見地からの整理・検討を行っていきたい．

謝辞

本報告書をまとめるにあたり，以下の方々に，メールあるいは面談の形で有益な情報をいただきました．お忙しい中，お時間を割いて丁寧に回答して下さいましたことに，ここに厚く御礼申し上げます．
桧野副部門長；新井副部門長；時間周波数科　今江科長；時間周波数科波長標準研究室　洪室長；時間周波数科周波数システム研究室　鈴山様，長さ計測科長さ標準研究室　平井様；力学計測科質量力標準研究室　上田室長；温度湿度科高温標準研究室　丹波室長，山澤様；温度湿度科放射温度標準研究室　石井室長，佐久間様；温度湿度科湿度標準研究室　北野室長；流量計測科　寺尾様；電磁気計測科電気標準第2究室　坂本様；電磁波計測科　小見山科長；電磁波計測科高周波標準研究室　島岡様，堀部様；電磁波計測科電磁界標準研究室　島田室長；先端材料科材料分析研究室　伊藤様；先端材料科高分子標準研究室　衣笠室長；計量標準システム科　前田科長，津越様；計量標準管理センター国際計量室　藤間室長；計量標準管理センター計量研修センター　福本様
有意義な調査研究のテーマを設定して下さった馬場物性統計科長，榎原応用統計研究室長にもここに御礼申し上げたいと思います．榎原室長にはメールでのインタビュー様式の作成，面談の日程調整などにもご尽力いただきました．また，平素より研究の相談にのってくださっている応用統計研究室の皆様にも厚く御礼申し上げます．

付録1　実際の問題と紹介した文献の対応

比較試験の実施後に起こりうる問題点と紹介した文献の対応を参考のために，4.3小節と5章中の文献の対応を表2としてまとめた．ごく簡単にまとめたものであり，

表2　4.3小節での分類と参考文献の対応

	4.3 小節での分類	文献の内容	参考文献 (備考)
試験前の検討	ガイドライン	CIPM ガイドライン文書	Guidelines for CIPM key comparisons [11]
	(a) 不確かさや補正量の算出	一般的不確かさの算出	GUM[68]
		試験所の報告値間の相関係数の計算方法	GUM F1.2.3，F.1.2.4
	(b) 同等性の評価，外れ値の検出	測定量が正規分布しないことが予想される場合	Steele ら[56]（拡張 χ^2 検定）
		最大数の試験所が一致性を確認し，KCRV の計算に参加するようにする方法	Cox [48]（LCS: Largest Consistent Subset）
	(c)複数被試験器や基幹比較間のリンク	同一比較試験内の複数試験器のリンク	公平に評価する統計的方法はまだ見ていない。ある一つのループに重みを置いてよい場合には比較試験間のリンク参照.
		比較試験間のリンク	Douglas ら[53]
	(d) 被試験器のドリフト，長期不安性，移送の不安定性	スター方式で行っている場合（長期不安定性の影響が各試験所の不確かさより大きい場合.）	Elster ら[62]（ただし，比較試験の最初と最後の試験所間の同等性の評価などには問題を残す.）
		ラウンドロビンの場合	下記の一致性が見られないときに，「未知で共通の不確かさがありそうで，考察に含めたい場合」が使えるかも知れない.
	(e) KCRV の計算	線形結合型の KCRV に対する考察	Kacker ら[52]
		実験の自由度を考慮した KCRV	Zhang ら[58]
	事前に比較試験で得られる報告値が従う確率分布に対する情報を持っている場合	ベイズ統計を用いた解析	Kacker ら[66], [67]
試験後の解析	一致性の検定通過しないときに不確かさや補正量の見積もりに考察を加えたい場合	未知で共通の不確かさがありそうで，考察に含めたい場合	Paul ら[63]
		実験方法の違いなどにより，未知の系統的誤差が生じ，その不確かさを見積もりたい場合	Willink[64]
		KCRV の値は妥当であるが，個々の報告した不確かさや補正量の見積もりが不十分そうである場合	Weise ら[65]

詳しい内容は本文中および紹介した文献を参照されたい．

付録2　外部標準偏差，内部標準偏差とBirgeレシオ

CIPMの基幹比較のガイドライン[12]に記載されていないので本文では省略したが，基幹比較でよく用いられるBirgeレシオ[69]について説明する．試験所間比較の枠組みの中で，時に外部標準偏差と呼ばれるのは，

$$u_{\mathrm{ext}}(y)=\sqrt{\frac{1}{N-1}\frac{\sum_{i}^{N}(x_i-y)^2/u^2(x_i)}{\sum_{i}^{N}1/u^2(x_i)}} \tag{A.1}$$

と表わされるものである．式の形を見てわかるように，これは重み付きの実験標準偏差に他ならない．

本文式(2.2)の$u^2(y)$は内部分散と呼ばれるもので，ここでは違いを明確にするために$u_{\mathrm{int}}^2(y)$と表記する．改めて，内部分散の正の平方根である内部標準偏差$u_{\mathrm{int}}(y)$を求める式に書き直すと，

$$u_{\mathrm{int}}(y)=\sqrt{\frac{1}{\sum_{i}^{N}1/u^2(x_i)}} \tag{A.2}$$

である．比較試験に一致性が見られるというのは，報告された不確かさから考えられる測定結果の標準偏差（内部標準偏差）が実際の測定結果のばらつき（外部標準偏差）に等しいということに他ならない．このため，この2つの標準偏差の比Rは一致性の指標として意味がある．

$$R=\frac{u_{\mathrm{ext}}(y)}{u_{\mathrm{int}}(y)}=\sqrt{\frac{\sum_{i}^{N}(x_i-y)^2/u^2(x_i)}{N-1}} \tag{A.3}$$

このRがBirgeレシオと呼ばれる．この式を本文式(2.3)と比較すると，

$$R=\sqrt{\frac{\chi_{obs}^{2}}{\nu}} \tag{A.4}$$

の形をしていることが分かるから，「Rが1であること」

表3　試験所間比較の用語に関する説明と本文中の記載箇所

	Term (English)	用語（日本語）	説明	本文中の記載箇所	出典
基幹比較関連	CIPM MRA (Mutual Recognition Arrangement)	計量標準の国際相互承認協定	国家計量標準と各国NMIが発行する校正証明書に関する相互承認協定	1章	2), 70)
	Key Comparison	基幹比較	CIPM MRAの枠組みで行われるCIPM CCおよびRMOが主体となって行う比較試験	1章	2)
	Supplementary Comparison	補完比較	CIPM MRAの枠組みで行われる2国間比較など基幹比較以外の国際比較	1章	2)
	CMC (Calibration and Measurement Capability)	校正・測定能力	通常の状態（normal condition）で顧客が得られる校正と測定の能力	1章	71)
	CIPM MRA Appendix B	計量標準の国際相互承認協定　附属書B	基幹比較および補完比較の結果	1章	2), 70)
	CIPM MRA Appendix C	計量標準の国際相互承認協定　附属書C	国家計量標準機関および指名計量標準機関のCMC	1章	2), 70)
	KCDB (Key Comparison Database)	基幹比較データベース	CIPM MRA附属書A-Dの4つの部分から成り立っているデータベース	1章	5)
	pilot　institute (pilot laboratory)	幹事機関	基幹比較の際に比較試験の調整者となる機関	2.1節	11)
	Draft A	ドラフトA	基幹比較の結果を最終報告する前に参加者のみに開示される暫定の報告書	2.1節	11)
	Draft B	ドラフトB	基幹比較の報告のためにCCに提出され，承認されれば，最終報告書となる文書	2.1節	11)
	KCRV (Key Comparison Reference Value)	基幹比較参照値*	基幹比較の測定量に対して，付与される合意値	2.1節，2.2節	11), 12)
	DOE (Degree of Equivalence)		基幹比較の結果として得られる同等性評価の指標	2.1節，2.2節	11), 12)
	Unilateral DOE		KCRVとの差とその不確かさ（95％信頼区間）で与えられるDOE	2.2節	2), 12)
	Bilateral DOE		各試験所間の差とその不確かさ（95％信頼区間）で与えられるDOE	2.2節	2), 12)
	χ^2 test	χ^2（カイ二乗）検定	一般的な一致性の検定方法であり、一致性の検定としてCIPM基幹比較のガイドラインで推奨される方法	2.2節	12)
	Procedure A	手順A*	CIPM 基幹比較のガイドラインで，χ^2 検定で一致性が確認された場合に推奨されるKCRVとDOEの計算方法	2.2節	12)
	Procedure B	手順B*	CIPM 基幹比較のガイドラインで，χ^2検定で一致性が確認されなかった場合に推奨されるKCRVとDOEの計算方法	2.2節	12)
技能試験関連	ISO/IEC Guide 43		試験所間比較による技能試験に関する国際規格（JIS Q 0043はこの翻訳規格）	1章，3章	6), 7), 14)
	ISO 13528		試験所間比較による技能試験のための統計的方法に関する国際規格（JIS Z 8405はその翻訳規格）	3章	13), 15)
	Coordinator	コーディネータ（など）	技能試験スキームの作業に必要な前活動を調整する責任を持つ組織（または者）	3.1節，3.2節	13), 15)
	Performance statistics	成績を表す統計指標（など）	試験所間比較の結果から，各試験所に与えられる成績の指標．統計的分析に基づく．	3章	13), 15)
	Assigned value	付与された値	成績を表す統計量の計算のために求められる試験所間比較において定められた測定量の代表的値	3章	13), 15)
	Standard deviation for proficiency assessment	技能評価のための標準偏差	使用可能な情報に基づく，技能評価に使用するばらつきの尺度	3.2節，3.3節	13), 15)
	Robust mean, Robust standard deviation	ロバストな平均，ロバストな標準偏差	JIS Z 8405附属書「ロバストな解析」に基づいて計算した平均値と標準偏差．それぞれ成績を表す統計指標の計算に用いられることがある．	3.3節	13), 15)
	z-score	zスコア	試験所のかたよりを標準化した成績を表す統計指標の一種	3.3節	13), 15)
	E_n number	E_n数	試験所のかたよりを標準化した成績を表す統計指標の一種	3.3節	13), 15)
	Youden plot	ユーデン・プロット	類似した2種の試験の実施結果（zスコア）をグラフィカルに表示するもの	3.3節	13), 15)

＊ これらの訳語はあまり一般的ではなく，日本語報告書などでも英語のまま表記されることが少なくない．

と，「χ_{obs}^2値がνであること」は同じ意味を持つ．χ^2分布の分散は2νであるが，これを分散2νの正規分布と置き換えて，$\sqrt{[1\pm2\sqrt{(2/\nu)}}$とか$1\pm\sqrt{(1/2\nu)}$の範囲にBirgeレシオが入っているかどうかで，$\chi^2$分布の表を使わない簡易な一致性の検定が行われることがある．

付録3　用語集

基幹比較およびそれ以外の試験所間比較に特有な用語は多くある．この文書ではなるべく多くの用語を取り上げて紹介したが，改めて関連する用語を，表3に一覧としてまとめた．分かりやすさを重視した解説を加えたが，必ずしも出典文書に示された公式な定義とは一致しないことがあるので，詳細に興味のある読者は出典元も参照されたい．

参考文献

1）　独立行政法人産業技術総合研究所：計量標準の国際相互承認，AIST TODAY, 4-2(2004)20-25.

2）　CIPM: Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued by National Metrology Institutes, Modified (BIPM, 2003).

3）　田中秀幸，トレーサビリティにおける校正方式の活用に関する調査研究，計量研究所報告 50, 181（2001）

4）　CIPM: Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued by National Metrology Institutes (BIPM, 1999).

5）　BIPM: The BIPM Key Comparison Database (KCDB), http://kcdb.bipm.org/.

6）　ISO/IEC Guide 43-1 Proficiency Testing by Interlaboratory Comparisons- Part 1: Development and Operation of Proficiency Testing Schemes, ISO/IEC (1997).

7）　ISO/IEC Guide 43-2 Proficiency Testing by Interlaboratory Comparisons- Part 2: Selection and Use of Proficiency Testing Schemes by Laboratory Accreditation Bodies, ISO/IEC (1997).

8）　ILAC: Use of Proficiency Testing as a Tool for Accreditation in Testing (ILAC, 2004).

9）　OIML: Framework for a Mutual Acceptance Arrangement on OIML Type Evaluations (MAA) - Amendment (2006), (OIML, 2006).

10）G. Guslicov and P. D. Bièvre: The 1st International Proficiency Testing Conference, Romania 2007, Accred. Qual. Assur. 13-2(2008)109-110.

11）CIPM: “Guidelines for CIPM Key Comparisons”, Modified, (BIPM, 2003)

12）M. G. Cox: The Evaluation of Key Comparison Data, Metrologia 39-6(2002)589-595.

13）ISO 13528 Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparisons, ISO (2005).

14）JIS Q 0043-1 試験所間比較による技能試験　第1部：技能試験スキームの開発及び運営，日本規格協会（1998）.

15）JIS Z 8405 試験所間比較による技能試験のための統計的方法，日本規格協会（2008）

16）D. W. Tholen: ISO/IEC 17043: The new International Standard for proficiency testing Accred. Qual. Assur. 13(2008)727-730

17）KCDB: CCT-K3 Summary results (BIPM, 2003).

18）B. W. Mangum, G. F. Strouse, W. F. Guthrie, R. Pello, M. Stock, E. Renaot, Y. Hermier, G. Bonnier, P. Marcarino, K. S. Gam, K. H. Kang, Y.-G. Kim, J. V. Nicholas, D. R. White, T. D. Dransfield, Y. Duan, Y. Qu, J. Connolly, R. L. Rusby, J. Gray, G. J. Sutton, D. I. Head, K. D. Hill, A. Steele, K. Nara, E. Tegeler, U. Noatsch, D. Heyer, B. Fellmuth, B. Thiele-Krivoj, S. Duris, A. I. Pokhodun, N. P. Moiseeva, A. G. Ivanova, M. J. de Groot and J. F. Dubbeldam: Summary of Comparison of Realizations of the ITS-90 over the range 83.8058 K to 933.473 K: CCT Key Comparison CCT-K3, Metrologia 39-2(2002)179-205.

19）KCDB: CCT-K5 Final Report (BIPM, 2008).

20）KCDB: CCT-K7 Final Report (BIPM, 2006).

21）M. Stock, S. Solve, D. del Campo, V. Chimenti, E. Méndez-Lango, H. Liedberg, P. P. M. Steur, P. Marcarino, R. Dematteis, E. Filipe, I. Lobo, K. H. Kang, K. S. Gam, Y.-G. Kim, E. Renaot, G. Bonnier, M. Valin, R. White, T. D. Dransfield, Y. Duan, Y. Xiaoke, G. Strouse, M. Ballico, D. Sukkar, M. Arai, A. Mans, M. de Groot, O. Kerkhof, R. Rusby, J. Gray, D. Head, K. Hill, E. Tegeler, U. Noatsch, S. Duris, H. Y. Kho, S. Ugur, A. Pokhodun and S. F. Gerasimov: Final Report on CCT-K7: Key Comparison of Water Triple Point Cells, Metrologia, 43-Tech. Suppl. (2006) 03001.

22）KCDB: CCL-K1 Final Report (BIPM, 2001).

23）R. Thalmann: CCL key comparison: calibration of gauge blocks by interferometry, Metrologia 39-2(2002)165-177.

24）KCDB: APMP.L-K1 Final Report (BIPM, 2005).

25）I. Fujima, A. Hirai, H. Matsumoto, N. Brown, S. Gao, R.

P. Singhal, C. -S. Kang, A. M. Dahlan, E. Howick, T. S. Leng, S. Charkkian and B. Q. Thu: APMP.L-K1: Calibration of gauge blocks by interferometry: Final Report, Metrologia, 43-Tech. Suppl.(2006)04004.

26）電磁波計測科 堀部氏私信

27）電磁波計測科 島岡氏私信

28）KCDB: CCEM-K8 Final Report (BIPM, 2003)

29）G. M. Reedtz and R. Cerri: Final Report on Key Comparison CCEM-K8 (dc voltage ratio), Metrologia 40-Tech. Suppl.(2003)01001.

30）G. M. Reedtz, R. Cerri, I. Blanc, O. Gunnarsson, J. Williams, F. Raso, K.-T. Kim, R. B. Frenkel, Z. Xiuzeng, A. S. Katkov, R. Dziuba, M. Parker, B. M. Wood, L. A. Christian, E. Tarnow, S. K. Mahajan, A. Singh, and Y. Skamoto: Comparison CCEM-K8 of DC Boltage Ratio: Results, IEEE Trans. Instrument. Meas. 52-2(2003) 419-423.

31）KCDB: CCM.FF-K1 Final Report (BIPM, 2007).

32）J. S. Paik, K.-B. Lee, P. Lau, R. Engel, A. Loza, Y. Terao and M. Reader-Harris: Final Report on CCM.FF-K1 for Water, Metrologia 44-Tech. Suppl.(2007)07005.

33）KCDB: CCM.FF-K3 Final Report (BIPM, 2007).

34）Y. Terao, M. van der Beek, T. T. Yeh and H. Müller: Final Report on the CIPM Air Speed Key Comparison (CCM.FF-K3), Metrologia, 44-Tech. Suppl.(2007) 07009.

35）KCDB: CCM-FF-K6 Final Report (BIPM, 2007).

36）J. Wright, B. Mikan, R. Paton, K. -A. Park, S. Nakao, K. Chahine and R. Arias: CIPM key comparison for low-pressure gas flow: CCM.FF-K6, Metrologia 44-Tech. Suppl.(2007)07008.

37）力学計測科 上田氏私信

38）先端材料科 衣笠氏私信

39）R. Nagahata, K. shirmada. K. Kishine, H. Sato, S. Matsutyama, H. Togashi and S. Kinugasa: Interlaboratory Comparison of Average Molecular Mass and Molecular Mass Distribution of a Polystyrene Reference Material Determined by MALDI-TOF Mass Spectrometry, International Journal of Mass Spectrometry 263-(2-3) (2007) 213-221.

40）C. M. Guttman, S. J. Wetzel, W. R. Blair, B. M. Franconi, J. E. Firard, R. J. Goldschmidt, W. E. Wallace and D. L. VanderHart: NIST-Sponsored Interlaboratory Comparison of Polystyrene Molecular Mass Distribution Obtained by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry: Statistical Analysis, Anal. Chem. 73-6(2001) 1252-1262.

41）高分子計測クラブ: MALDI-TOFMS 共同測定2007-1結果報告（高分子計測クラブ，2007）.

42）K. Ito, T. Oka, Y. Kobayashi, Y. Shirai, K. Wada, M. Matsumoto, M. Fujinami, T. Hirade, Y. Honda, H. Hosomi, Y. Nagai, K. Inoue, H. Saito, K. Sakaki, K. Sato, A. Shimazu, and A. Uedono: Interlaboratory comparison of positron annihilation lifetime measurements for synthetic fused silica and plycarbonate J. Appl. Phys. 104-2(2008)026102.

43）独立行政法人 産業技術総合研究所，青森県工業総合研究センター：知的基盤部会分析分科会 平成19年度 第50回分析技術共同研究；第39回分析技術討論会総合資料（2007）（内部資料）

44）社団法人 日本環境測定分析協会: ISO/IECガイド43-1に基づく技能試験（Webサイト），https://prc.jemca.or.jp

45）高橋千晴，北野寛，横田冨夫: JCSS湿度技能試験としての持ち回り測定，第62回応用物理学会講演会（応用物理学会，愛知工業大学，2001年9月）.

46）H. Preston-Thomas, P. Bloembergen and T. J. Quinn: Supplementary Information for the International Temperature Scale of 1990 (BIPM, 1997).

47）独立行政法人産業技術総合研究所 計量標準総合センター訳編集：国際文書第8版（2006）/日本語版 国際単位系（SI） 安心・安全を支える世界共通のものさし，（日本規格協会，2007）.

48）M. G. Cox: The Evaluation of Key Comparison Data: Determining the Largest Consistent Subset, Metrologia 44-3(2007)187-200.

49）J. E. Decker, A. G. Steele and R. J. Douglas: Measurement science and the linking of CIPM and regional key comparisons, Metrologia 45-2(2008)223-232.

50）近藤良夫，舟阪渡: 技術者のための統計的方法，（共立出版，1967）.

51）日本分析化学会: 日本分析化学会（Webサイト），http://www.jsac.or.jp/

52）R. N. Kacker, R. U. Datla and A. C. Parr: Statistical analysis of CIPM key comparisons based on the ISO Guide, Metrologia 41-4(2004) 340-352.

53）R. J. Douglas and A. G. Steele: Pair-Difference Chi-squared Statistics for Key Comparions, Metrologia 43-1(2006) 89-97.

54）R. Willink: On the Interpretation and Analysis of a Degree-of-Equivalence, Metrologia 40-1(2003)9-17.

55）M. G. Cox and C. Eiø: The Generalized Weighted Mean of Correlated Quantities, Metrologia 43-4(2006) S268-S275.
56）A. G. Steele and R. J. Douglas: Extending Chi-Squared Statistics for Key Comparisons in Metrology, J. Comp. Appl. Math. 192-1(2006)51-58.
57）H. K. Iyer, C. M. Wang and D. F. Vecchia: Consistency Tests for Key Comparison Data, Metrologia 41-4(2004)223-230.
58）N. F. Zhang: The Uncertainty Associated with the Weighted Mean of Measurement Data, Metrologia 43-3(2006) 195-204.
59）R. Willink: Forming a Comparison Reference Value from Different Distributions of Belief, Metrologia 43-1(2006)12-20.
60）C. M. Sutton: Analysis and Linking of International Measurement Comparisons, Metrologia 41-4(2004)272-277.
61）N. F. Zhang: H.-K. Liu, N. Sedransk and W. E. Starawderman: Statistical Analysis of Key Comparisons with Linear Trends, Metrologia 41-4(2004) 231-237.
62）C. Elster, W. Wöger and M. Cox: Analysis of Key Comparison Data: Unstable Travelling Standards, Measurement Techniques 48-9(2005)883-893.
63）R. C. Paule and J. Mandel: Consensus Values and Weighting Factors, J. Res. Nat. Bur. Stand., 87-5(1982)377-385.
64）R. Willink: Statistical determination of a comparison reference value using hidden errors, Metrologia 39-4(2002) 343-353.
65）K. Weise and W. Wöger: Removing model and data non-conformity in measurement evaluation, Meas. Sci. Technol. 11-12(2000)1649-1658.
66）R. N. Kacker, A. Forbes, T. Kessel and K.-D. Sommer: Classical and Baysian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluations, Metrologia 45-3(2008)257-264.
67）R. N. Kacker, A. Forbes, R. Kessel and K.-D. Sommer: Baysian posterior predictive p-value of statistical consistency in interlabotatory evaluations, Metrologia 45-5(2008)512-523.
68）BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML: Guide to the Expression of Uncertainty in Measurement, 2nd ed. (ISO, 1995).
69）R. T. Birge: The Calculation of Errors by the Method of Least Squares, Phys. Rev. 40-2(1932)207-227.
70）EURAMET編、産業技術総合研究所　計量標準総合センター、製品評価技術基盤機構　認定センター訳編、“計量学- 早わかり METROLOGY - IN SHORT”, 第3版, 2009.
71）CIPM: Calibration and Measurement Capabilities in the context of the CIPM MRA (CIPM, 2008).