寄　稿

試験における不確かさ評価について

独立行政法人　産業技術総合研究所
計測標準研究部門応用統計研究室長　榎原研正

1. はじめに

測定における不確かさの表現のガイド[1]（GUMと略称）が出版されておよそ10年がたち，「不確かさ」という言葉は計測や試験の分野において広く知られるところとなっている。ISO/IEC 17025は，試験所に「測定の不確かさを推定する手順を持ち，適用する」ことを要求しており，試験関係者の不確かさに対する十分な技術的理解は急務になっている。

GUMが提示した不確かさの概念は，GUM以前に問題となっていた，測定結果の誤差評価や精度表示の方法の不統一や混乱を解消するうえで有効であり，計量学的にも優れたものであった。しかし，様々の技術分野でそれぞれの目的のもとで実際に不確かさ評価を行うためには，GUMの一般的記述の理解だけでは必ずしも十分ではないため，不確かさ評価の現場で新たな混乱が生じていることも事実である。ここでは，試験における測定の不確かさ評価においてしばしば議論の対象となる問題の幾つかについて，GUMの立場に忠実に，しかしあまり肩肘張らずに，多少の交通整理を試みたい。

2. 不確かさの定義をめぐって

GUMによる不確かさの定義は，それが回りくどい表現に見えることもあって，十分な注意が払われていないことが多いように思われる。しかしそのもってまわった表現が，不確かさ評価の技術論においても本質的に重要と思われるので，あらためてここで議論しておきたい。

不確かさは次のように定義されている。

不確かさ＝測定の結果に付随した，合理的に測定量に結びつけられ得る値*）のばらつきを特徴づけるパラメータ

ポイントは，「合理的に測定量に結びつけられ得る値（the values that could reasonably be attributed to the measurand）」とは何かということである。「合理的に」や「結びつけられ得る」という表現は必ずしもわかり易い言葉ではないが，「それが測定量の値であると主張しても，不合理とは言えないような値」，あるいはもう少しわかりやすくは「測定量の真値**）の候補」と言うことができる。

この定義が意味することは，不確かさの評価にあたっては，真値の候補と考えられるような値の集合をまず想定しなさい，ということである。そしてその集合の拡がりの大きさ，具体的には標準偏差を推定することができれば，それが標準不確かさに他ならないということになる。このような集合は，ある値からある値までという明確な境界

*）下線は筆者

**）GUMでは，「測定量の真値」というときの「真」という修飾語は冗長であるために用いないという立場をとっているが，ここでは分かり易さに配慮して真値という言い方を使うことにする。試験における真値とは何かというやや哲学的問題についても触れない。

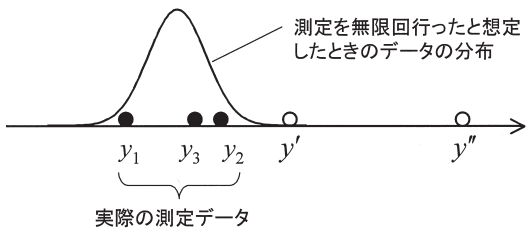

図1　合理的に測定量に結びつけられ得る値

をもつものではない。例えば，ある試験片の質量を3回測定して，**図1**に示す3個のデータy_1，y_2，y_3を得たとしよう。y_1からy_2の間にある一連の値はすべて真値の候補と言って差し支えないだろう。それだけでなく，この範囲から少しずれた例えばy'も，可能性は低くなるが，候補であると言っても不合理とはいえない。しかし，これらからずっとはなれたy''は，候補とは考えられないか，少なくとも真値である可能性はずっと低いと考えるのが合理的であろう。

具体的に真値の候補とは，例えば，測定を無限回行ったときのデータの集合を想定すれば良い。それは例えば**図1**に示すような分布（確率分布）で表されるものである。このような確率分布で規定されるような値の集合，それが「合理的に測定量に結びつけられ得る値」である。集合の境界はぼやっとしており，その中には真値である可能性が高いものも低いものもある。この事情をGUMでは次のように説明している（**付属書D5.2**）。

「ある与えられた測定量及びその測定結果に対して，一つの値があるのではなくて，あらゆる観測やデータおよび物理的世界についての知識と矛盾しないような，いろいろな可能性の大きさをもって測定量に結びつけられ得る無限個の値がある*)」

「いろいろな可能性の大きさをもって」というのは確率分布で規定されるということを意味する。また，「あらゆる観測や…」は，このような値の集合として，真値の可能性のあるものは全て入れなさい，ということを意味している。この点は特に重要である。例えば，短期間の繰り返し試験のばらつきよりも，理由は不明であるけれども，日をかえた試験のばらつきの方が大きいことはしばしば経験する。どの日に行った試験が特に正しい，という根拠がないならば，異なる日のデータはすべてこのような集合のメンバーとして考えなければならない。同様のことは，異なる試験者によるデータ，異なる試験機を使って得たデータ，さらには異なる試験所で得られたデータについて言える。もっとも，不確かさを評価するときには，必ず異なる試験日，試験機，試験所などによるばらつきを求めなさい，ということではない。これについては，以下の「**5.** 原因追求型評価と原因不問型評価」で議論する。

以上のようなGUMの定義に忠実な不確かさの評価とは，次のようなものと考えられる。

1）試験によって得られる可能性のある値の集合をまず想定しなさい。その値としては，現実に目の前にある試験システムによって得られるものだけでなく，規定に矛盾しない範囲で試験を実施したときに生じ得るあらゆる値を含むこと。

2）その値の中には，真値である可能性が高いものも低いものも含まれる。可能性の大小を表す確率分布を何らかの方法で推定しなさい。

3）その分布の拡がりを標準偏差で表しなさい。それが標準不確かさである。

実際に最終の測定結果yの不確かさをこの方法で求めるのは，確率分布が複雑になりすぎて難しいことが多いので，この方法が適用可能な成分まで分解して評価しておき，あとでそれらを伝播則を使って合成しようというのが，不確かさの合成の考え方ということになる。

3. 何の不確かさ？

*）訳は筆者。

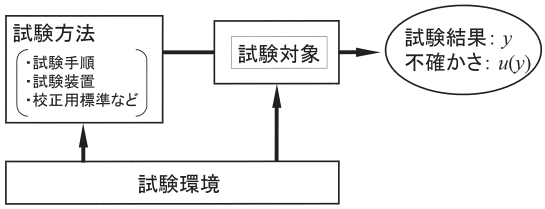

図2　試験の方法，対象，環境が特定されないと不確かさは決まらない

GUMでは元来，不確かさを，特定の「測定結果」の信頼性の尺度を表すものと規定している。定義で「測定の結果に付随した」としているのも，また標準不確かさを常に$u(y)$の形（値yの不確かさ）で表記しているのも，この点を明確にする意図があると考えられる。すなわち，「値」の信頼性であって，試験方法や試験そのものといった「方法」や「作業」の信頼性の指標と考えるのは，拡大解釈である。細かい点に見えるが，この点を忘れた不確かさ評価はしばしば混乱する。

しかし現実には不確かさ評価を個々の試験の都度行うのは困難なので，このような拡大解釈はいろいろの局面で行わざるを得ない。**ISO/IEC 17025**に基づく試験所認定制度も，このような拡大解釈を前提に成立していると考えられる。

図2に示すように，試験結果 y とその不確かさ $u(y)$は，試験方法，試験対象，試験環境のすべてが特定されて初めて決まる。例えば質量測定では，対象が1 kgなのか10 kgなのかで，$u(y)$も異なり得る。もし$u(y)$が成分に分解できて，例えば二乗和の形で，

と分解できるならば，この内の$u_{方法}$をもって，対

$$u^2(y) = u^2_{方法} + u^2_{対象} + u^2_{環境} \quad \cdots (1)$$

象や環境を特定しない一般的な「試験方法の不確かさ」という言い方が許されるかも知れないが，現実には上式のような加法性が期待できないことは明らかである。

従って，ある「試験サービス」の代表的不確かさを評価したいという局面であっても，試験対象と試験環境を特定せざるを得ない。試験環境は自然に特定されることが多いと考えられるので，特に意識する必要があるのは代表性のある試験対象の特定，つまりどのような値をもつ対象を幾種類選定すればよいか，という問題に帰着する。

この問題について現時点で合意された解答はない。測定量Yの大きさとして例えば1から100（単位は任意）にわたる2桁の範囲での試験を行っているならば，およその値が$Y_1=3$，$Y_2=10$，$Y_3=30$，$Y_4=100$と，ほぼ等比数列をなす4種類の試験対象を選定するのが，一つの考え方であろう。4つの範囲1〜3，3〜10，10〜30，30〜100のそれぞれにおける不確かさを，Y_1，Y_2，Y_3，Y_4に対して評価した不確かさで代表するとすれば良い。各範囲の上限値を選ぶのは，測定量が大きくなるほど一般に不確かさも大きくなることと，試験依頼者に対して保証できる不確かさ（不確かさはこれを超えないという値）を提示するために，不確かさの上限を求めることが妥当*）と考えられることによる。不確かさの測定量Yへの依存性が重要でないならば，$Y_1=10$，$Y_2=100$の2種類でも構わない。試験範囲がもう少し狭い場合は，等差数列のように選ぶのが自然な場合もあろう。たとえば測定量が40から100の範囲をカバーするならば，$Y_1=60$，$Y_2=80$，$Y_3=100$とする。

もう一つの考え方として，範囲毎に不確かさを示すのでなく，不確かさを測定量の大きさYの関数として示すやり方が有用と考えられる。例えばデジタル電圧計などの性能表示で，正確さなどをなどと表記（digitは最小表示桁，readingは表示

*）これは不確かさは一般に過大評価するのが良いということではない。不確かさが過大評価されていると，それを引用するBタイプ不確かさが適切に行えない。

$$2\ \text{digits} + 0.0001 \times \text{reading} \quad \cdots (2)$$

値）することが昔から行われているが，これと同様の考え方である。Y_iとしては，全試験範囲で代表的な3ないし5点程度を選ぶことで実用的に十分であることが多いだろう。最終的には，不確かさを測定量の大きさに対して回帰分析し，何らかの単純な関数形で表現する。

GUMの拡大解釈をさらに進めて，不確かさを「測定器」や「試験装置」の属性のように扱うのは適当でないことを特に指摘しておきたい。測定器が決まっても，それをどのような標準を使って校正するか，校正頻度はどうするのか，繰り返し何回の平均値を測定結果にするのかなどの測定手順，さらに測定対象や環境にも不確かさは依存するので，これらを特定しない「測定器の不確かさ」というものはあり得ないからである。試験装置や測定器の性能の指標としては，繰り返し性，再現性，直線性，分解能など従来からの指標が依然として適当なのであって，不確かさはこれらを置き換えるものではない。

4. 校正用標準がある場合とない場合

不確かさそのものは，すでに第2節で述べたように，「真値の候補の拡がりの程度」なので，試験の不確かさでも測定（あるいは校正）の不確かさでも，概念として違うということはない。

ただし，評価という点からは

測定：校正のための標準が利用可能

試験：標準が利用可能でない

という違い（必ずしも一般則ではないが）による，多少系統的な違いはあり得るであろう。標準が利用可能で，例えば1週間に一度の周期で校正を行っている測定の場合，とりあげるべき不確かさ成分は，次の3成分に帰着する。

(1) 標準の値がもつ不確かさ（校正証明書から評価。Bタイプ評価）

(2) 校正する際の不確かさ（標準をn回繰り返し測定した平均値を校正で用いるとして，$s/\sqrt{\ }$。ただしsは繰り返し測定のばらつきの標準偏差。）

(3) 校正後の不確かさ＝1週間の中で生じ得る測定のばらつき（時間的変動プラス繰り返し測定のばらつき。）

この内（**3**）は，例えば，同一対象を月曜日から金曜日について午前・午後の2回，計10回測定したときの標準偏差を求めればよい。この中には時間的変動と繰り返しばらつきの両方が入っている。もし繰り返しのばらつきと時間的変動を分離して求めたければ，測定の各々で繰り返し2回のデータを取っておき，合計20個のデータを分散分析すればよい。これで求まる繰り返し測定のばらつきは，（**2**）でも使える*)。結局，不確かさ評価の中心は（**3**）のばらつきの評価となるので，標準が使える場合，不確かさ評価は一般に素朴である。

一方，標準が利用可能でない試験の場合，試験結果の確からしさは，データの目に見えるばらつきの大きさではなく，規定されている試験条件と現実の試験条件の乖離の程度が問題となる。これには系統的なずれも含まれるので，標準がある場合よりはやっかいである。試験結果に影響するすべてのパラメータについてこれを評価するために，評価は一般に複雑で本格的なものとなる。これについては，次項で触れる。

GUMにおける不確かさ評価の骨格は，測定量

*）厳密にいうと，標準を測る場合と実際の測定物を測る場合で，繰り返しのばらつきが有意に異なることがあり，この場合は，(2) と (3) での繰り返しのばらつきは別の値を使う必要がある。例えば，測長において，標準は金属のブロックゲージを使うが，実際の測定対象物は測定力で変形し得るプラスチック，などの場合である。

5. 原因追求型評価と原因不問型評価
—ボトムアップとトップダウン —

y を入力量 x_i の関数として

と表しておき，y の不確かさを x_i の不確かさを

$$y = f(x_1, x_2, ..., x_n) \quad \cdots\cdots\cdots (3)$$

合成して求めるという点にある。このためには，y を変動させる主要な原因がすべてわかっていることが前提になる。これは原因追求型の不確かさ評価と呼べる。各原因の不確かさを積み上げるので，ボトムアップ方式と言える。この場合，入力量となり得るのは，試験片の幅，負荷などの試験条件を表すパラメータか，温度，湿度などの環境条件を表すパラメータであり，評価時に人為的に制御可能な変数（いわゆる母数型因子）である。原因追求型では，感度係数（$\partial f/\partial x_i$）と，入力量の標準不確かさ $u(x_i)$ が別々に評価される。規格に基づく試験は，試験条件や環境条件のパラメータを一定の値（あるいは範囲）に制御することにより試験結果が一意的に決まるという前提で成立しているので，ボトムアップ方式の評価は自然な方法と言える。

一方，試験者，試験機，日の違いなどによって測定結果が変動し得るが，その変動の真の物理的原因が何か追求しない，あるいは追求してもわからないという立場での評価があり得る。これは原因不問型の評価といえる。実際には，例えば試験機を替えることによって式（3）におけるそれぞれの x_i が変動し，これが y の変動となっているのだが，物理的因果関係には関心を払わずに，試験機が替わったときに y がどれだけ変動するか，だけに着目した評価である。変動の要因として取り上げるのは，いわゆる変量型因子（例えば試験機を例にとると，その背後に無数の試験機を想定し，実験対象とする数台の試験機は，その中からランダムに選ばれたものとして取り扱うことができる

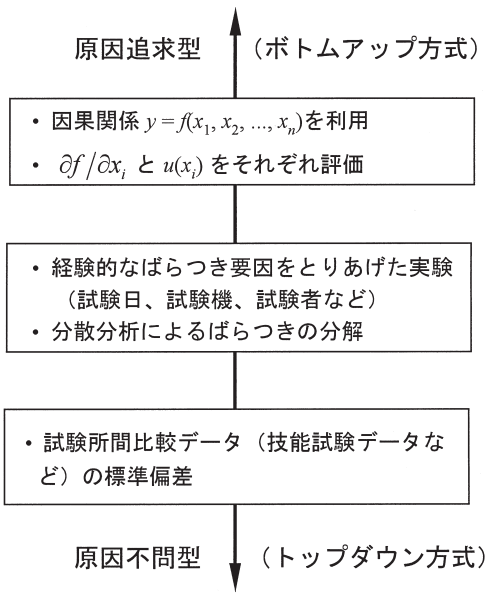

図3　原因追求型評価と原因不問型評価

因子）である。試験日，試験者などもこの種の因子の例である。「合理的に測定量に結びつけられ得る値」の全域をカバーするためには，できるだけ様々な因子を取り上げるのが望ましいので，この方式では，統計的な実験計画にもとづく実験と分散分析の利用が有効となる。

一方，技能試験のように多数の試験所が参加した比較試験データがあるときには，これを利用した原因不問型評価を行うことができる。試験所間のばらつきには，試験条件のばらつきや環境条件のばらつき，試験者による違い等の主要なばらつき要因の多くが含まれると考えられるので，不確かさの評価のためには適当なデータといえる。ばらつきの要因毎の分解も考えないので，上に述べた評価方法より，さらに「原因不問」に徹した方法といえる。この方法では試験所間に差がつかないが，試験所の技術的能力の高低が重要でない試験では，とりわけ有効な評価方法と考えられる。最終的な測定量についての変動を直接評価するの

で，トップダウン方式と言える。

以上を**図3**に整理する。ボトムアップ方式だけがGUMに準拠した方法である，ということではない。トップダウン方式の数学モデルは，入力量xが1つしかなく，しかも入力量がそのまま測定量yであるという特別なケースに相当する。

また，原因不問型で使う分散分析の手法は，

$$y = x \quad \cdots\cdots (4)$$

GUMにおいてもその重要性が指摘されている（**4.2.8**，**付属書H5**など）。

6. 破壊試験の不確かさ評価

破壊試験では，複数の試料に対して得た試験データのばらつきが，試験方法の不完全さからくるのか，試料の特性のばらつきからくるのか，判断できない点が大きな問題となる。ボトムアップ方式の評価を利用することで，この問題を回避した不確かさ評価ができるだろうか。例えば，コンクリートの圧縮強度試験では，円柱型の試験用試料について，試料の断面積 A（mm²）と，破壊するまでの最大荷重（N）から，圧縮強度 F（N/mm²）が次で計算される。

この式は，式（3）の測定の数学モデルの一例

$$F = \frac{P}{A} \quad \cdots\cdots (5)$$

である。上の問題の回避には，複数の試験用試料に対してFを求め，そのばらつきを計算するということをせずに，不確かさ評価する必要がある。このため，Aについてはノギスによる円柱試料の断面積測定の不確かさ$u(A)$，Pについては荷重測定の不確かさ$u(P)$を求め，これらを合成するので良いだろうか。$u(P)$は単なる荷重測定としての不確かさなので，試験の物理プロセスに伴う不確かさがこれでは評価できていないと考えられる。

この事情は，例えば，炭化部分の面積から繊維の難燃性を試験する場合（**JIS L 1091**）に，布地の面積測定の不確かさだけ考慮しても，試験全体の不確かさ評価になっていないのと同様である。同一加熱条件下で生じ得る炭化の進行度合いのばらつき，ある部分が炭化したかどうかの判定のばらつきなどが，試験の不確かさの本質的部分と考えられる。

従って，コンクリート圧縮強度の場合でも，圧縮による破壊という物理プロセスに付随する不確かさを評価する必要がある。そのため，入手可能な最良の技術を用いて可能な限りばらつきの少ない試験用試料を作成し，これらに対してもなおかつ試験結果がばらつくならば，それは試験の不確かさとして評価しよう，というのが破壊試験の場合の基本的考え方である。このばらつきには，実際には試料の特性のばらつきによるものも含まれるのだが，分離しては評価できないので，試験の不確かさに含めるのである。過大評価にならざるを得ないのだが，過大部分をできるだけ小さくしようという考え方と言える。

破壊試験の不確かさ評価で特徴的なこととして，ばらつき成分とかたより成分を分けて考え，ばらつき成分は，最終の測定量に対して評価し（トップダウン型），かたより成分は個別の入力量について評価したものを合成する（ボトムアップ型）のが自然である。コンクリートの圧縮強度試験の例でいうと，ばらつきはFのばらつきとして，かたよりはPやAについての評価を合成する。これは，PやAについて個別のばらつきの評価を行うと，それらの間の相関を考慮しなければならず（Aが大きい試料はPも大きいことが期待されるので）これを避けたいこと，およびかたより成分は個別の入力量について評価する以外に方法がないことによる。実際にどのように評価を進めるべきかについては，本誌の中の上園氏による解説を参考にされたい。

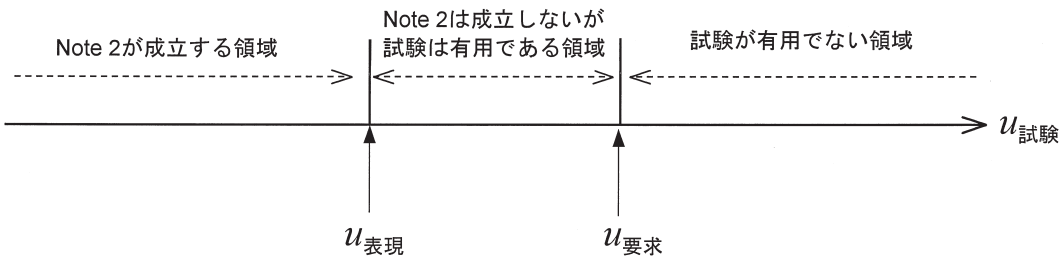

図4　$u_{試験}$ はどこにある？

7. ISO/IEC 17025，5.4.6.2 Note 2について

すべての試験の不確かさが正確にわかっているというのは理想的な状況であるが，不確かさ評価にはコストがかかるので，無駄な評価を行わないことは，試験所にとっても社会全体にとっても重要である。次の**ISO/IEC 17025**の**5.4.6.2**項**Note 2**は，この点に配慮したものと考えられる。

「広く認められた試験方法が測定の不確かさの主要な要因の値に限界を定め，計算結果の表現形式を規定している場合には，試験所はその試験方法及び報告方法の指示に従うことによって，この項目（試験所が測定の不確かさ評価を行うという要求*）を満足すると考えられる。」

しかし，**Note 2**が技術的に本当に妥当なのかどうかは必ずしも明白ではなく，この点は現在の論点の一つになっている。

Note 2の「計算結果の表現形式を規定している」典型的な例としては，表記される試験結果の有効桁数が規定されている場合があげられるであろう。例えば**JIS A 1108**「コンクリートの圧縮強度試験方法」は，供試体の圧縮強度，及び見掛け密度をともに有効数字3桁に丸めるよう規定している。圧縮強度の試験結果が，例えば70.1 N/mm^{2}と表示される場合，これは70.05 N/mm^{2}から70.15 N/mm^{2}の範囲内にあるということを意味するので，この丸めに伴う標準不確かさは

この大きさを，表現形式の規定で決まる不確かさ

$$\frac{0.05}{\sqrt{3}} = 0.029 \quad (\mathrm{N/mm^2}) \qquad \cdots\cdots\cdots (6)$$

として $u_{表現}$ と表すことにしよう。一方，試験依頼者が要求する不確かさを，標準不確かさの上限値（これ以上になると困るという値）として $u_{要求}$，また試験所が行う試験の実際の標準不確かさを $u_{試験}$ としよう。

試験結果が依頼者にとって有用である必要条件はである。この条件の成立は，試験にとって本質的

$$u_{試験} \leq u_{要求} \qquad \cdots\cdots\cdots (7)$$

に重要である。一方，依頼者は，広く認められた試験方法で結果がどのように表現されるか認識していると考えられるので，
は自動的に成立していると期待できる。従ってもし

$$u_{表現} \leq u_{要求} \qquad \cdots\cdots\cdots (8)$$

が成立しているならば，(7) の条件は自動的に満

$$u_{試験} \leq u_{表現} \qquad \cdots\cdots\cdots (9)$$

足されることになる（**図4**）。

幾つかの広く認められた試験について，式 (9) が実際に当てはまるかどうかを検証することは必要であろう。その内の幾つかについては (9) がおそらく実際に成立する。しかし，有効数字はし

*）括弧内は筆者

ばしば最小のばらつき成分（繰り返しばらつきや分解能など）を考慮して決められるから，現在の「広く認められた試験方法」がすべて式（9）を満足するということは期待できないと思われる。

Note 2をより寛容に解釈して，広く認められた試験方法では，仮に（9）が成立していなくとも，（7）は成立しているものと考える（その試験の有用性が認知されているということは，すでに（7）が成立していることを暗に示していると考える）のか，もしくは**Note 2**の妥当性の検証がもっと厳密に求められるようになるのか，あるいは他に選択肢があるのか，現時点では明らかでない。今後検討を深めつつ，国際的にも受け入れられるような解を捜していく必要がある。長い目で見た場合，**Note 2**の解釈論以外に，今後作成あるいは修正される試験規格の中に，不確かさの考え方をどう取り入れて行くかの検討を進めることが重要である。

不確かさに関わる現在の混乱の解消のためには，独りよがりな考え方に陥らず，広く合意が得られるような現実的で妥当な解を捜していくことが何より重要である。GUMをバイブルのように扱ってその解釈論のみを展開することは適当ではないが，最終的な拠り所はGUMしかないので，GUMに十分則った議論が望まれる。本稿では，しばしばGUMにおいて読み落とされがちな点を含めて，試験分野の不確かさ評価で重要と思われるポイントの幾つかについて検討を加えた。不確かさ評価におけるより現場的な問題点の検討についてはここでは触れることができなかった。これについては文献2）を参照して頂きたい。

【参考文献】

1）Guide to the Expression of Uncertainty in Measurement（ISO，初版1993，修正版1995），日本語訳：計測における不確かさの表現のガイド（日本規格協会，1996）．

2）榎原，試験・分析における不確かさ評価の問題点と対策，計量管理，51，16-24（2002）

プロフィール

榎原研正（えはらけんせい）

（独）産業技術総合研究所
計測標準研究部門
応用統計研究室長

□**学歴・学位**
京都大学理学部卒・工学博士

□**専門・研究テーマ**
計測に関わる応用統計技術，
エアロゾル計測

8. おわりに