

会報

2015・4
第 70 号

Japan Association of Reference Materials

目次

- 1. 化学計測に役立つ『データの見える化』 1
- 2. 標準物質協議会見学会報告 9
- 3. 編集後記 12

化学計測に役立つ『データの見える化』

国立研究開発法人 産業技術総合研究所
物質計測標準研究部門 計量標準基盤研究グループ
城野克広

0. はじめに

産業技術総合研究所 城野克広と申します。不確かさ評価や試験所間比較データの解析をはじめとする測定データの統計的処理を専門に、研究開発・技術相談を行っております。日ごろのお仕事の中でお悩みのことがあれば、是非 k.shirono@aist.go.jp までご相談下さい。

まず、今回の連載のテーマをざっと説明したいと思います。右の数字をじっくりと見て下さい。そして、そこからどんなことが言えそうか考えてみて下さい。

次に、次ページの図1を見てみて下さい。

x	20.0	25.9	29.5	29.5	25.9
y	30.0	28.1	23.1	16.9	11.9

x	20.0	14.1	10.5	10.5	14.2
y	10.0	11.9	16.9	23.1	28.1

この図から、どんなことが言えるでしょうか？実は図1は、上の表の数字を散布図にしたものです。つまり、表と図の持っている情報はほとんど同じです。

しかし、数字だけを見て、「データが円周上の点になっている」と分かったという方は、なかなかいないと思います。一方、グラフを見れば、誰でも、これは円周を等分に分割した点を表していると分かります。

この例で、データが円周上の点だと言いたいのであれば、データを数字として見せるのではなく、図として見せた方がはるかに効率的です。誤解を恐れずに言えば、数字というのは、それ自体には価値はありません。その数字がどういう意味を持っているのかという情報に価値があるわけです。つまり、その情報をどう伝えるかということが、大事なわけです。

この連載では、化学計測で役立つそうなくつかのグラフの読み描きについて、お話したいと思っています。これを通して、どのようにデータに意味を持たせていくかのヒントになればと考えています。中身は決して簡単ではないかも知れませんが、気楽に読んでいただけるように、化学計測に限らず、様々な題材を選んで、説明していきたいと思っています。

1. ヒストグラムの見方と見せ方

1. 1 ヒストグラムの見方

最初にみなさんと考えていきたいのは、図2のようなグラフです。ちょっと、嫌な(?) 感じのデータではありますが、上のようなグラフから何が分かるか、そして、それをどのように描くかが最初のテーマです。

ちなみにこれは、厚生労働省が実施した平成21年の国民生活基礎調査の概況の所得の分布状況として報告されたものです(参考文献1)。ぱっと見て、どういうグラフかは伝わるとは思いますが、一応、丁寧に見ていきましょう。横軸は世帯所得金額について、一番最後を除いて、100万円ごとに範囲を定めて、示したものです。最後の級の金額の幅は無限大です。この範囲のことをビンと呼びましょう。縦軸は、全体の世

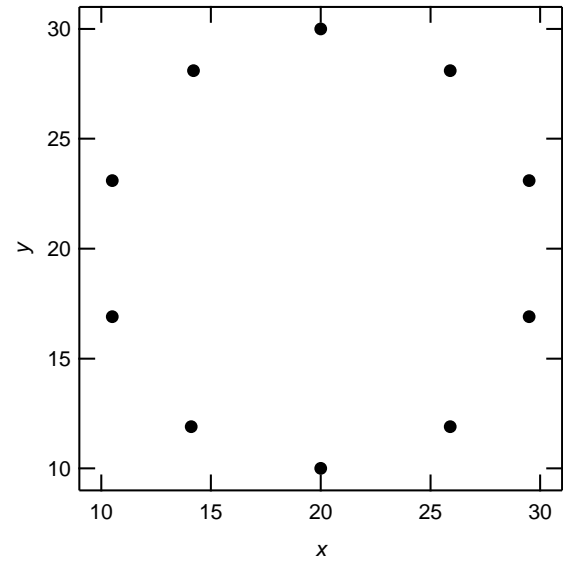


図1 散布図

帯の中で、そのビンに属する世帯がどのくらいの割合であるかを示しています。このように横軸を調べた値(測定値)の範囲、縦軸をその範囲に当てはまる値の割合で示した棒グラフのことをヒストグラムと言います。縦軸は、このように割合にすることもありますが、実際の値の数(上の場合なら世帯数)とすることもあります。

ヒストグラムを用いると、測定された値がどのように分布しているのかよく分かります。少しグラフを眺めていただいて、「世帯所得というのは、全体としては、どういうばらつきを持っているか?」という質問に対する答えを考えていただけますでしょうか。例えば、「2000万円以上稼いでいる世帯が1.2%ある。」確かに、上のヒストグラフからも分かります。しかし、それは別にヒストグラムにしなくても分かったはずで、ヒストグラムにしないと分からないこととは何でしょうか?

ヒストグラムを見る際に、気にしていただきたいのは、「単峰性(ピークがただ一つであるか)」と「対称性(ピークを中心に右側と左側の測定値の数が同じくらいか)」です。

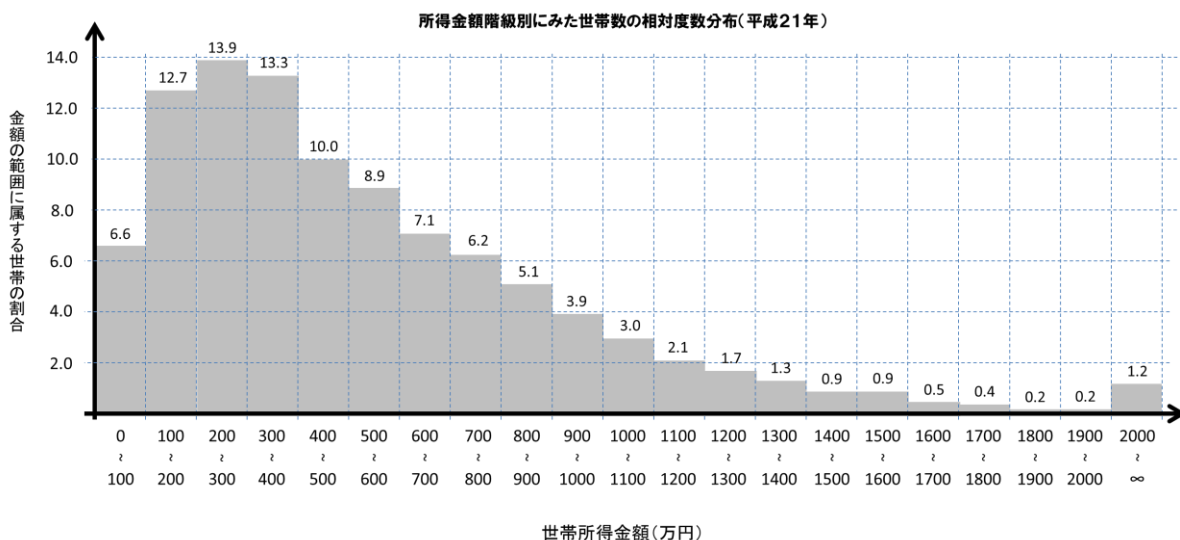


図2 所得金額階級別に見た世帯数の相対度数分布(平成21年) [参考文献1]

世帯所得の分布をみていると、単峰ではありそうです。先ほども述べた通り、200~300万円のあたりにただ一つのピークがあります。左右対称であると言われると、全くそうではありません。ピークを基準とすると、明らかに右側にたくさんの世帯数が所属しているように見えます。ここでは、「世帯所得の分布は、全体としては単峰であり、そのピークよりも大きい所得を持つ世帯が多い」という印象をざっと持っていたら、ヒストグラムから得られる重要な情報を押さえていると言えると思います。

計測の分野においても、測定値の分布が「単峰であるかどうか」ということはとても重要です。もし、ヒストグラムが2つないし3つのピークを持っているとしたら、そのデータを何も考えずに処理してしまうのは大変危険です。ピークがたくさんあるということは、すなわち、測定が失敗していると考えた方が無難でしょう。例えば、2人の測定者がおり、それぞれ高めと低めに値を出してしまうために、2つのピークがあるのだとすれば、可能な限り2人の値が接近するような教育・訓練を実施してから測定プロセスを始めるべきです。2つのピークの生じる原因ごとに、データを層別する必要がある場合もあるでしょう。

さて、「対称であるかどうか」もヒストグラムから読み取れるはずの重要な情報なのです。しかし、実のところ、その判断は難しいものです。次ページの図3のAとBのヒストグラムを見て下さい。いずれも100個のデータからヒストグラムを作ったものです。どちらもおよそ単峰に見ることができそうです。AのデータとBのデータのどちらが左右対称と言えるでしょう？

実は、この2つのグラフは、全く同一の(しかも左右対称な)データ群から、ランダムに100個ずつをとった値です。ということは、もっとたくさんの測定を行えば、この2つは完全に同一の(左右対称な)ヒストグラムにはならないのです。Aの方が非対称で、Bの方が対称に見えると言いたいところですが、単にデータの個数が少なかったため、偏って見えたということなのです。数限りなく多くのデータがあれば、より確定的なことが言えます。しかし、化学計測では、100個もの測定値を同一の環境下で取得するのは難しいと思います。結論として、化学計測の実務では、データの対称性の判断を、ヒストグラムを使って直観的に行うのは、困難と言ってよいと思います。

というわけで、ヒストグラムから分かる重要なことは、「分布のピークが一つしかない」とい

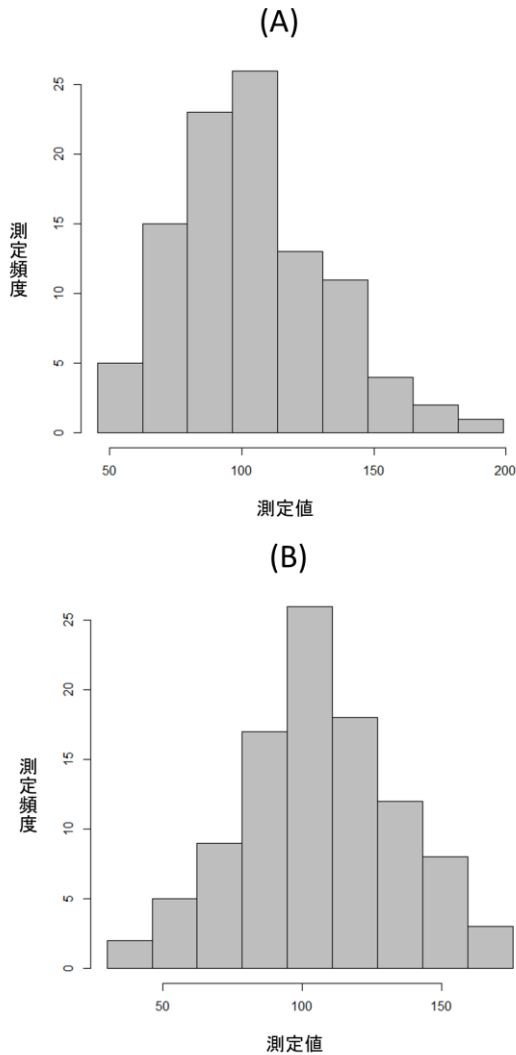


図3 どっちが左右対称？

うことです。これにより、データが、一定の環境下で取られた揃ったものであることを確かめるといわけです。

1. 2 ヒストグラムの見せ方

ここまでの内容を踏まえまして、図4のA~Cのデータは、それぞれ30個の測定値について、ヒストグラムを描いたものです。このうち、測定値が100~120あたりに、唯一のピークを持つような分布を持つのはどれでしょうか？

答えはCのデータ…だけではなく、全部です。実は上の3つのヒストグラムはすべて完全に同一のデータから描いたものです。図3とは異

なり、完全に同じ数字のデータから違うヒストグラムを描いています。しかし、Aのヒストグラムを見れば、3~4つのピークが確認でき、このデータには単峰性はないように見えます。Bのヒストグラムを見ると、単峰ではありそうですが、そのピークは60~100あたりにありそうです。

ヒストグラムというのは、分布をみるために便利な道具ですが、使い方を誤ると誤った結論を導きやすいものなのです。ヒストグラムを描くとき、以下の3つのポイントに気をつけるとよいのではないかと思います。

- ① ビンを細かくしすぎない。
- ② ビンの幅は基本的には等しくする。
- ③ ビンの数を奇数にする。

ひとつめのポイントはよく言われることです。具体的にビンの数をどれほどにしたらよいとよく言われるかという、データ数 N に対して、 \sqrt{N} くらいにするとよいと言われます。今回の場合には $N=30$ ですから、 $\sqrt{30}=5.477\dots$ となるため、5ビンか6ビンくらいがよいということになります。

2つ目のポイントは、単峰性を確認するという目的のためには欠かせない条件です。図4の(B)のヒストグラムの60~100のビンにみるように、幅を大きくすると、測定頻度もそれに従って大きくなってしまいます。ただし、所得分布のヒストグラムの2000万円以上のように、それ以上あるいはそれ以下の範囲に興味がない場合に限り、最後のビンの幅を極めて大きくすることはあり得るでしょう。これ以外の理由で、幅の大きさを違うものにするのは、ミスリードを導きやすく大変危険です。

3つめのポイントは、私の個人的な趣向かも知れませんが、多くの場合ヒストグラムは左右対称に近い形になりますから、もし、単峰であるとする、そのピークは真ん中に来る可能性が高いわけです。

しかし、ビン数が偶数だと、ちょうど真ん中はビンとビンの境なので、最大のビンは、真ん中より左側か右側に来てしまいます。単峰性を確認するという目的を考えれば、興味ある範囲でビン数を奇数にすることは勧められるのではないかと考えています。

図4のCのヒストグラムは、測定値のうち最小の値と最大の値で与えられる範囲を $\sqrt{30}$ を超えない最大の奇数である5で等分して、ヒストグラムを描きました。いかがでしょうか？この手法がいつもベストということはないかも知れませんが、とっかかりを得るためのひとつのヒントになればと思います。

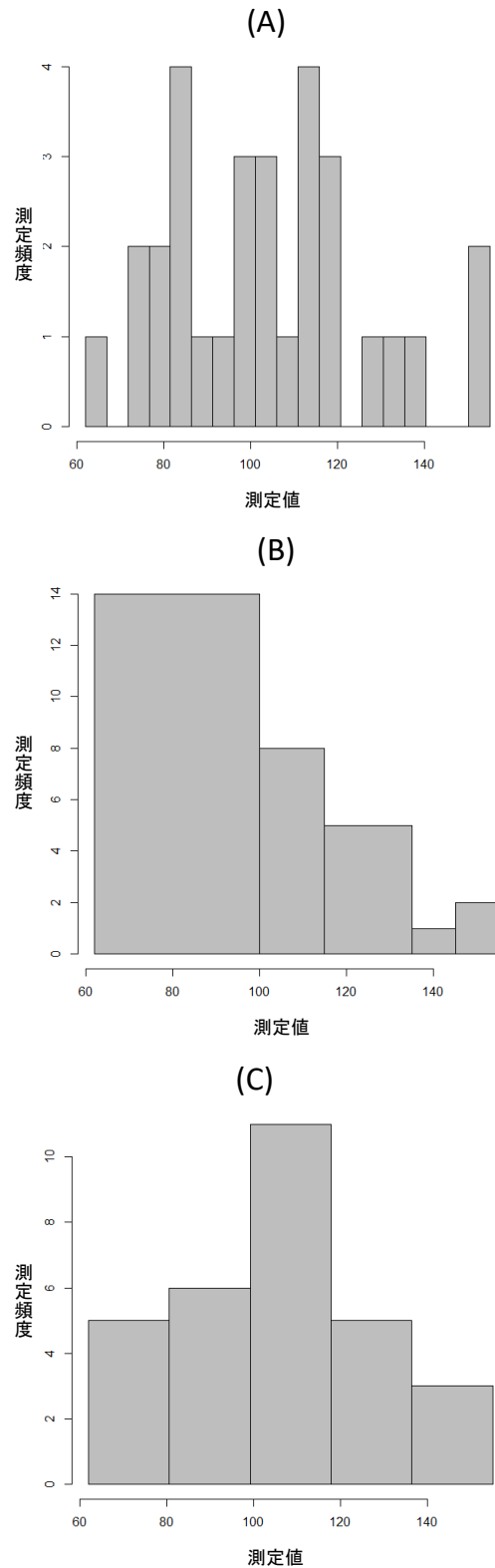


図4 良いヒストグラムとは？

2. エラーバーの見方と見せ方

2. 1 エラーバーの見方

図5は、○が測定値、そして、棒がそのばらつきの範囲を表したグラフです。この**ばらつきの範囲を表す棒のことをエラーバー**と言います。では、エラーバーの表す「ばらつきの範囲」とは一体何でしょう？

実はこれは、正解のない質問です。色々な分野でよく使われてきたこのエラーバーですが、人によって使い方が異なるために、図5だけを見て、エラーバーが何を意味するかは、一概には言うことはできません。

通常、グラフには説明書きがついています。

(図5では省略しましたが。) 例えば、「○○を測定した結果。ただし、エラーバーは**標準偏差**を示す。」というの是一個の典型でしょう。あるいは「標準偏差」の代わりに「標準偏差の2倍」というのもしばしば使われます^{注1)}。

標準偏差とは何かということを確認しておきましょう。ある一連のデータについて、図6のようにヒストグラムを描けたとします。図の①と②のうち、標準偏差に当たるのはどちらでしょう？言うなれば、①はデータ全体のばらつきを考えると、このくらいずれるのは普通と考えられる程度のずれ、②はこんなにずれることは少ないかなという大きなずれです。

正解は…①です。標準偏差とは普通 (= 標準) のずれ (= 偏差) という意味です。ちなみに、②はちょうど①の倍の大きさになっております。標準偏差の2倍というのは、実は通常のデータのばらつきを考えると、こんなにずれることは少ないかなという大きなずれに当たります。このために、エラーバーの大きさとして標準偏差の2倍が使用されることも多いのです。エラーバーがある場合、きちんと説明を読まないと、その意味を誤解してしまうことになりかねません。特に、それが「普通のずれ」なのか、「大きなずれ」なのかということをよく見ておく必要があります。

2. 2 エラーバーの見せ方

エラーバーをどう見せるかと言っても、ヒストグラムのような細かい話はなく、標準偏差をどのように計算するかということにつきます。ただ、これが決して簡単というわけでもないのです。というのは、同じデータから得られる標準偏差には、ざっと言って以下の3通りのものがあるからです。

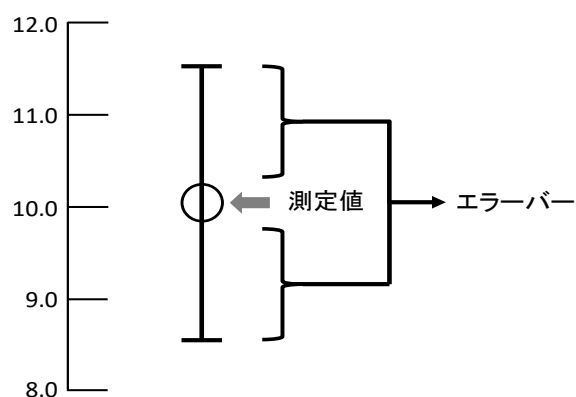


図5 エラーバーとは

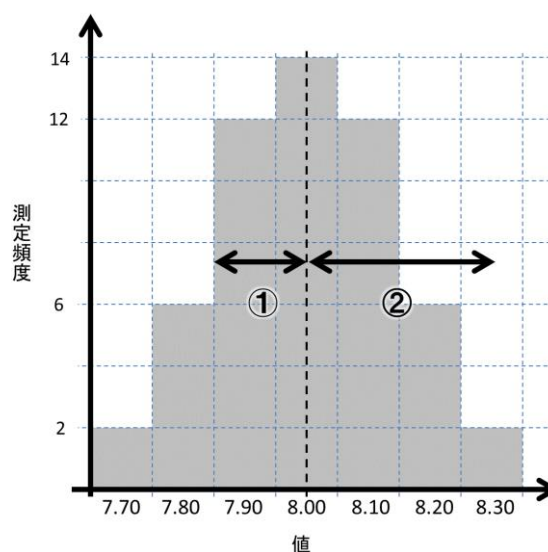


図6 標準偏差は①と②のどっち？

- (1) ひとつの値のばらつきの標準偏差
- (2) 平均値のばらつきの標準偏差
- (3) 予測値のばらつきの標準偏差

$$s_{\text{mean}} = \frac{s}{\sqrt{5}}$$

ひとつの値のばらつきの標準偏差から、説明していきましょう。例えばある池の水に含まれる物質 Z の濃度について、5回サンプリングした結果として、以下の x_1 、 x_2 、 x_3 、 x_4 、 x_5 という濃度値を得たものとします。これらの平均値を \bar{x} としています。

ある池の水を5回サンプリングし求めた物質 Z の濃度値 (mg/g)

x_1	x_2	x_3	x_4	x_5	\bar{x}
4.5	5.2	4.8	5.0	5.5	5.0

標準偏差は平均値からどのくらいずれるかという指標ですから、 $(x_1 - \bar{x}) \sim (x_5 - \bar{x})$ がどのくらいの大きさを調べる必要があります。一般的な統計学の教科書には、標準偏差 s は以下のように計算すると書いてあると思います。

$$s = \sqrt{\frac{1}{5-1} \left\{ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_5 - \bar{x})^2 \right\}}$$

ひっかかるところがあるとすれば、「なぜ2乗するのか」ということと、「なぜ5ではなく(5-1)で割るのか」というところかと思いますが、今回は割愛させていただきます^{注2)}。

化学計測の分野では、この標準偏差を用いる機会は少ないと思います。なぜならば、この標準偏差は測定された x_1 、 x_2 、 x_3 、 x_4 、 x_5 が \bar{x} に比べてどのくらいばらついたかという指標ですが、報告するのは \bar{x} だからです。1回測定した値よりも、5回測定した値の平均値の方が質の良い値ということに異論をはさむ方はいらっしゃらないと思います。「質の良い」とは、統計学的観点から言えば、ばらつきが小さいという意味です。

平均値のばらつきの標準偏差とは、この \bar{x} のばらつきのことです。具体的には、平均値の標準偏差 s_{mean} は以下の式で与えられます。

分母のルートの中には、繰り返し数である5が入っており、 s_{mean} は s より小さい値となります。これは、その池の全体的な平均としての物質 Z の濃度 \bar{x} を知りたいとき、その値に付すエラーバーのための標準偏差として役立つものです。

ただ、もしかすると、平均値ではなく、その池からコップ一杯をサンプリングして、そのサンプリングされた水中の物質 Z の濃度に興味があるかも知れません。そのコップの水を飲むとすれば、平均的には大丈夫でも、そのコップにはより濃く物質 Z が入っていて、体に害があるかも知れません。もちろん、飲む水を測ることはできません。コップの水については、どのように考えたらよいのでしょうか？

予測値のばらつきの標準偏差はそのようなときに役立つ考え方です。つまり、その飲み水を仮に測定したとしたらどんな値になるかということです。予測値は \bar{x} となります。しかし、そのばらつきは、平均値のばらつきよりも大きくないと不自然でしょう。実のところ、予測値の標準偏差 s_{rep} は、 s_{mean} より大きく、

$$s_{\text{rep}} = \sqrt{s^2 + s_{\text{mean}}^2}$$

と与えることとなります^{注3)}。

さて、図7にそれぞれ、

- (1) ひとつの値のばらつきの標準偏差
- (2) 平均値のばらつきの標準偏差
- (3) 予測値のばらつきの標準偏差

の2倍を標準偏差として、エラーバーを描きました。さきほども申し上げましたように、(1)は実際には、使われることはないと思います。しかし、(3)と誤って使われることがまれにあります。(1)と(3)の値はほとんど同じになることが多いですが、それでも正しい値が用いられているかは、よく考えていただきたいところです。

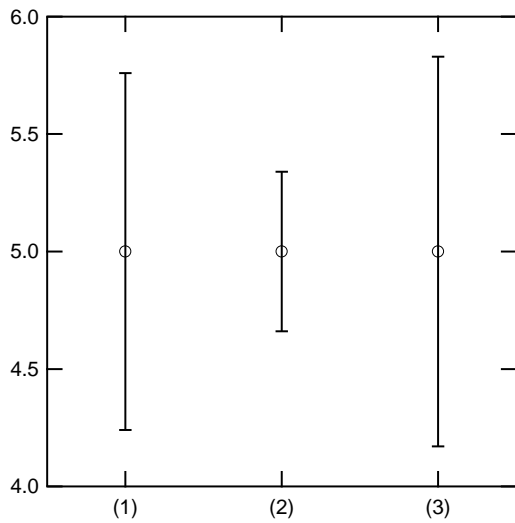


図7 (2×標準偏差)で描いた(1)ひとつの値、
(2)平均値、(3)予測値のエラーバー。

注1) 図の説明に「エラーバーは95%の信頼区間を示す。」ということもしばしばありますが、95%信頼区間のおおざっぱな意味合いは標準偏差の2倍と似たようなものです。信頼区間に興味のある方は、参考文献2 p.65「3.2 正規分布を用いた母平均の区間推定について」を参照して下さい。

注2) 標準偏差の計算の仕方に興味のある方は参考文献2 p.28「2.6 母分散と標本分散」を参照して下さい。

注3) 注意して欲しいのは、これはあくまで将来測定したときの値の予測値であって、将来の測定の操作のばらつきも含んでいることです。この測定の操作のばらつきが分かる場合にはそれを抜いたものが、コップに含まれている物質Zの濃度のばらつきです。このようなケースに興味のある方は参考文献2 p.103「4.7 標準物質の値付けとそのばらつきの大きさの推定」を参照して下さい。

参考文献

1. 厚生労働省が実施した平成21年 国民生活基礎調査の概況、<http://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa09/2-2.html> (2015/03/11 最終確認)
2. 田中秀幸著、「分析・測定データの統計処理ー分析化学データの扱い方」、朝倉書店

★ 次回は要因効果図とパレート図というグラフを紹介したいと思います。

標準物質協議会見学会報告

一般財団法人化学物質評価研究機構
化学標準部 技術第二課
沢田貴史

平成27年3月4日（水）に標準物質協議会主催の見学会が開催されました。今回の見学機関は、警察庁の附属機関の一つである科学警察研究所（科警研）ということで、なかなか普段はお目にかかれない警察関連施設の内部を見学させていただきましたので、情報漏洩による業務執行妨害とならない程度にご報告したいと思います。



当日は、つくばエクスプレス「柏の葉キャンパス」駅に集合し、春の訪れを感じさせる陽気に誘われ、徒歩にて目的地へ向かいました。歩くことおよそ15分、到着したのはセキュリティ対策のためか、しっかりと門の閉ざされた7階建ての立派な建物でした。受付を済ませ、会議室にて研究所のパンフレットと見学スケジュールを受け取り、見学会企画をご担当いただきました瀬戸法科学第三部長、鈴木附属鑑定所長、井上法科学第三部付主任研究官より概要を

ご説明いただきました。科警研では、科学捜査についての研究や実験、これらを活用した鑑定・検査・犯罪の防止等、多岐にわたる科学捜査が行われています。近年、事件や事故の原因の一つにもなっている危険ドラッグ（旧称：脱法ドラッグ、違法ドラッグ）についての研究も盛んに行われているようです。科警研と名称のよく似た機関として、某TVドラマでもお馴染みの科捜研（科学捜査研究所）がありますが、こちらは各都道府県警に置かれた組織であり、国家機関である科警研とは管轄が異なります。科警研の内部組織は「総務部」「法科学部（生物学、工学、化学、情報科学等）」、「犯罪行動科学部」、「交通科学部」等に分かれており、今回は標準物質と比較的関連の深い「法科学第三部（化学）」の研究室及び附属鑑定所を見学させていただけるということでした。

次に、科警研全般に関するビデオを拝聴しました。沿革や組織説明の後、各部門紹介の際には、犯罪現場（恐らく本物）の写真から、サンプル採取の流れ、どの様に被害者の本人確認を行うか等、比較のスリリングな内容で、思わず見入ってしまいました。



その後、A班とB班に分かれて所内の見学に移りました。それぞれの班の見学スケジュールを見たところ「15:11 ○○研究室→15:21 ○○研究室→・・・」のように、なんと1分単位でおよそ2時間もの見学コースが決められているのには驚きました。ここからは、各研究室で見学した内容をいくつか紹介します。

まず、化学兵器関連の研究室では、硫化水素やフッ化水素、アルシン等の毒性ガスを用いることがあるため、アルカリ溶液洗浄のスクラパー付ドラフトにより外部への漏洩を防止していました。さらに、入退室の際には一旦小部屋を通る仕組みで、内部の物質が入退室時に漏洩しないよう小部屋はどちらか一方のドアしか開かないように工夫されており、扱っている物質の特性上、万が一、ということのないよう十分に配慮されていました。

また、実際の研究内容について具体的なお話も伺うことができました。以前、硫化水素を用いた自殺が多発した際に、事故現場に残された硫化水素ガスの発生するバケツ等の処理について多くの問い合わせがあったそうです。その対応について研究したところ、石灰を混ぜて内容物を塩基性にすれば反応が止まると結論付けたということでした。この様に対応に急を要する場合は、身近なものや容易に入手できるものを用いる必要があり、単純に購入した試薬等で対応を考えればよいという訳ではないことに、科警研での研究の難しさを感じました。

毒物関連の研究室では、標品とする試薬は表示値を鵜呑みにせず、自分たちで純度や不純物を確認することから始まり、天然毒については山へ採りに行くこともあるとご説明いただきました。標準物質を扱うものとしては表示値を信用していただけないのは複雑な心境でしたが、物質によってはかなり成分が変化するものもあるということでした。標準物質として用いられるような試薬に関する情報として、試薬開封後に通常の保存条件の他、多少ラフな保存条

件でどの程度変化するか等の定量的な情報をもっと開示できれば、一般的なユーザーも標準物質の安定性や使用期限等について、より留意するようになる可能性もあるかと思います。

麻薬や覚せい剤等の薬物関連の研究室では、各自の試薬(麻薬等)は金庫に入れられており、他の担当者が扱っているものには触れられないよう厳重に保管されていました。また、薬物使用の被疑者の頭髪分析においては、クリーンな部屋で実施しないと裁判の際にコンタミネーションの可能性が指摘されるとのことで、スペース的にクリーンな部屋を確保するのが難しい各地の科捜研から分析を依頼されることもあるとのことでした。

附属鑑定所では、偽造硬貨の判別等を行っており、特に偽造の多い五百円硬貨について、顕微鏡による拡大写真を用いながら、どこで偽造か否かを判断するかという点や、新五百円硬貨に盛り込まれた偽造防止のポイント等を紹介していただきました。特に新五百円硬貨の側面に刻まれた斜めのギザは、プレス加工により作製した硬貨を金型から容易に取り出すことができなくなるため、偽造が極めて難しいとのことでした。見学者のうち、数名の方が実際にご自分の五百円硬貨を凝視されていたようですが、もちろん偽造硬貨ではありませんでしたよね？

今回は、多種多様な化学関連の研究室を見学させていただき、この場で全てを紹介できないのが非常に心苦しいですが、誌面の都合上、割愛させていただきます。上述のとおり、分刻み



のスケジュールには驚きましたが、実際の見学もほぼ時間通りに終了したことに驚嘆いたしました。時間的にも肉体的にもかなりハードな見学コースだったものの、それ以上に研究内容が興味深かったためか、あっという間の2時間だったように感じました。

最後に、このように貴重な機会を提供していただき、また、非常にわかりやすくご説明いただきました科学警察研究所の関係者の方々にこの場をお借りして厚く御礼申し上げます。

編集後記

桜の季節も過ぎ暑さを感じる日もありますが、皆さまいかがお過ごしでしょうか。

会報第70号をお届けいたします。

城野様には、測定値から何が見えるかについて、何回かの連載でご紹介いただくこととなっています。私たちが測定結果を正しく利用するための基礎的な情報になると思います。測定値の信頼性に加え、得られた結果を正しく理解することが非常に重要であると思いますので、連載を楽しみにしたいと思います。城野様のご所属は、原稿をいただいた時点では独立行政法人でしたが、この4月から国立研究開発法人に名



(写真は、満開が少し過ぎたころの埼玉県杉戸町の桜です。)

称変更がありましたので、新しい名称としました。

沢田様には、科学警察研究所の見学報告を書いていただきました。千葉会長にアレンジいただき、多くの方に見学会にご参加いただきました。普段はなかなか訪問できない研究所だと思います。貴重な経験となりました。

会報に対するご意見等いただければと思います。また、今後とも皆様からのご寄稿をいただきたく、よろしくお願い申し上げます。

(四角目)

〒345-0043

埼玉県北葛飾郡杉戸町下高野 1600 番地

一般財団法人化学物質評価研究機構内

標準物質協議会 事務局 四角目和広

Tel. 0480-37-2601 Fax. 0480-37-2521

E-mail shikakume-kazuhiro@ceri.jp