

データのばらつきに向き合う

—ある国際協力での評価活動を振り返って—

日本原子力研究開発機構

原子力人材育成センター

原田 秀郎

harada.hideo@jaea.go.jp

1. はじめに

核データの精度を高めるには、どうすればよいか。実験屋であれば、測定技術を高めることにより精度向上を目指す。これは正攻法である。ただし、精度が向上していくと、無視できない補正因子が多くなるだけでなく、各補正因子の精度向上も必要になる。測定技術だけでなく、測定データの解析技術も高度化が必要になる。

核データの精度を高めるために、もう一つのアプローチがある。こちらが本稿の本題である。異なる実験チームの測定データを複数持ち寄り、適切に処理して、最適値とその不確かさを求めるアプローチである。ある統計学の式によると、独立したデータが4つ集まれば、2倍の精度向上が期待できる。測定技術の高精度化のアプローチでは、2倍の精度向上に少なくとも5年を要する[1]という経験則を考えると、その魅力は大きい。ところが、このように集めたデータには、各測定データの不確かさを超えたばらつきが、しばしばある。ばらつきのあるデータから、どのようにして最適値とその不確かさを求めればよいのであろうか？そのために適用できる統計学の数式はあるのであろうか？このような問題に向き合うこととなったある国際協力活動について、ばらつくデータの対処法に焦点を当てて、振り返ってみたい。

ある国際協力とは、国際核データ評価ワーキンググループ (WPEC) の41番目のサブグループ (SG-41) のことである。昨年7月に報告書[2]が公刊された。本活動は、 ^{241}Am の熱中性子捕獲断面積に焦点を絞り、これに詳しい世界中の専門家が、情報や知見を持ち寄って徹底的に検討を行うことにより、国際的に合意できる最適値とその不確かさ(本WGとしての推奨値)を導出しようとしたものである。次章で記載するいきさつから、著

者が主査を務めることとなった。

推奨値の導出に必要と考えられた微分実験、積分実験、崩壊データ、核データ評価という 4 つの専門技術を網羅できるよう、参加した専門家は、総勢 18 名となった。内訳は、日本 5 名、フランス 4 名、ベルギー 2 名、ドイツ 2 名、スペイン 2 名、米国 1 名、ロシア 1 名、中国 1 名である。微分実験、積分実験には様々な手法があり、そのための特殊な施設も必要となる。1 国で全ての種類の施設を有するのは難しくなっている。最近の測定を全て網羅し、詳細に議論できるのは国際協力のメリットである。参加した専門家の多くは、本データの測定に直接取り組んだ実験屋である。少しばかり頑固なところはあるものの、最適な推奨値を得るといった価値を共有し、真摯に突っ込んだ議論を行えるメンバーが揃った。さて、このようなメンバーで、測定値に相互に矛盾がある場合、国際的に合意できる推奨値にたどり着けるのか、いったいどのようなことになるのか、皆様は本稿を読む前に、どのように予想されるであろうか？

2. SG-41 活動のきっかけ

本活動の様子を紹介する前に、活動のきっかけを少し記載しておきたい。

テーマとなった ^{241}Am の中性子捕獲断面積は、核変換システムの設計などで高精度化が求められる最重要核データの一つである。10 年くらい前より、J-PARC の ANNRI を用いた測定プロジェクトや CERN の n_TOF プロジェクトなど、世界的に測定研究が活性化していた。マンハッタンの Sheraton New York Hotel & Towers で開催された核データ国際会議 ND2013 では、 ^{241}Am の中性子捕獲断面積測定についての進捗が、多くの実験チームから報告された。著作権の扱いが厳しくなっているためか、あるいは解析途中であったためか、口頭で報告された内容は会議録の内容を遥かに超えていた。これは会議に参加して、直接話を聴かないと分からないことである。

晚餐会がないかわりにコーヒブレークの充実が記憶に残る本会議で、 ^{241}Am の中性子捕獲断面積について、Peter Schillebeecx 氏、Matthias Rossbach 氏と自然と議論が始まった。原子炉中性子を用いた測定手法と TOF 法を用いた測定手法に関する比較検討を協力して進めると有益であろうことがすぐにわかった。このコーヒブレークでの立ち話が、本サブグループを提案するきっかけとなった。

核変換用核データ研究活性化の背景を遡れば、J-PARC での ANNRI プロジェクトが始まる前に、オメガ計画、加藤敏郎氏や故関根俊明氏らと著者との共同研究、小林捷平氏らの研究、井頭政之氏らの研究、石川眞氏が活躍された WPEC の SG-26（主査は故 Massimo Salvatores 氏）活動を始め、我が国の先駆的取組が数多くあるが、ここでは割愛させていただきたい。

3. SG-41 活動の様子

SG-41 の会合は、パリにある NEA 本部の会議室で 3 回行われた。初めに、本 SG-41 開始前の評価済み核データについて、岩本修氏によるレビューがあり、実験屋がその現状を共有した。次に、測定に携わった研究者から、各実験チームが行った実験と解析について、詳しい説明が行われた。各報告に対し、通常の国際会議では考えられないほど多くの、そして遠慮のない質疑応答が行われた。これは大変有益で、論文をいくら読み込んでも残っていた疑問について、論文の著者自身から話を聴くことで合点することが多くあった。例えば、2つの校正手法が論文に記載されており、どうして精度の低い校正手法を採用したのか疑問に残っていたことが、実験中のミスにより精度の高い方は採用できなかったことなどの裏話を聴いて氷解していった。このような検討にかなり時間を要したものの、各測定データが徹底的に吟味されていった。自らの測定あるいはデータ解析に問題が残っていること、そしてすぐには補正が困難であることを認識した研究者は、自ら進んで潔く推奨値への反映を辞退した。このようにして、推奨値の導出に用いるべきデータが絞り込まれていったのだが、中には難しいこともあった。

研究者の出身は、物理系であったり、化学系であったりする。このようにバックグラウンドが異なると、同じ核データを導出したとはいえ、持っている専門知識も異なる。分野の異なる研究者が導出したデータが食い違う場合、議論がかみ合わなかったりして、データの取捨選択について意見の収束が困難となることもあった。先ほども書いたが、参加者は自分で合点しないと譲れない頑固な人、もとい各専門分野の碩学である。数日の会議期間中に収束させることができなかったことは、宿題となる。意見の異なる 2 人の碩学を仮に A と B とする。帰国後、主査の役目として、A と B とに個別に連絡を取り、双方の理解と協力を得て、収束を目指した。「協力を得て」とさらっと書いたが、新たな補正を加えたデータの再解析など、大変骨の折れる協力を頂くこととなった。真摯な研究者との交流は、ありがたく、心に残る。もう一つ、「個別に連絡を取り」と書いたが、連絡する前には両方の考え方をよくよく理解しておく必要があり、雑事と相まって連絡が遅くなってしまった。これが報告書完成を遅くした主要因の一つである。最終的に、補正が必要であることが指摘されたデータについて、再解析して補正が行われたものについては、採用すべきデータとして残すことを基本としつつ、補正が期間中にできないものや不確かさの大きなものは除外することとした。

このようにして、補正すべきデータの補正と、採用すべきデータの絞り込みが進んだ。つぎ込んだエフォートは大きいものの、個人で文献を調査し、評価するのではこのように深い検討は難しいのではないだろうか。

4. ばらつくデータの扱いについて

次のステップとして、絞り込んだデータから推奨値を導出する必要がある。SG-41 で採

用した導出法を説明する前に、複数の測定結果から推奨値を導出する考え方について、いくつか紹介したい。

Norman Barford 氏（酒井英行氏訳）の参考書 [3] に、「矛盾している実験結果」という項目があり、次の指摘がある：「一方または両方の実験に系統誤差があるのではないかと疑うならば、とるべき最良のものは、実験を詳細に調べ、前もってこのような誤差を防ぐにはどのような用心をすべきだったかを理解しておくことであろう。それでもなお疑いが残るなら、(中略) もっと測定をするように頼むことである！」。この参考書では、数理統計学的なデータの処理方法が色々と説明されているが、最後にこの指摘が記載されており、矛盾するデータの扱いの難しさが窺える。

核データの精度向上の困難について 1979 年当時の様子がわかる記事 [4] を寄稿された吉澤康和氏に、系統誤差の扱いに詳しい「新しい誤差論」という貴重な本 [5] がある。この第 6 章 6 節に「評価の方法」という解説があり、そこで「外部誤差」と「内部誤差」に関し、「評価値の誤差として内部誤差と外部誤差の大きい方を採用している人が多い。」との記載を見ることができる。「評価値の誤差として内部誤差と外部誤差の大きい方を採用」という考え方は、Said Mughabghab 氏も Atras [6] で採用している。この手法の考え方は明確で、式も直感的に理解しやすい。後でもう少し詳しく述べるが、SG-41 ではこの手法を採用することとした。本手法を以下に記載しておく。

【SG-41 で採用した推奨値およびその不確かさの導出方法】

採用すべきデータの絞り込みの結果、 n 個の独立な測定値がある場合、その重み付き平均値 \bar{A} を、以下の式で計算し、これを推奨値として採用する。

$$\bar{A} = \frac{\sum_{i=1}^n \omega_i A_i}{\sum_{i=1}^n \omega_i} \quad (1)$$

ここで A_i は各測定値であり、 ω_i は各測定値の不確かさの逆数の 2 乗 ($= 1/\delta A_i^2$) である。重み付き平均値 \bar{A} の不確かさ $\delta\bar{A}$ は、内的不確かさ(δ_{int})と外的不確かさ(δ_{ext})の大きな方を採用する。ここで、内的不確かさ(δ_{int})および外的不確かさ(δ_{ext})は、それぞれ以下で定義される。

$$\delta_{\text{int}} = \frac{1}{\sqrt{\sum_{i=1}^n \omega_i}} \quad (2)$$

$$\delta_{\text{ext}} = \frac{\sqrt{\sum_{i=1}^n \omega_i (A_i - \bar{A})^2}}{\sqrt{(n-1) \sum_{i=1}^n \omega_i}} \quad (3)$$

この(1)式は、複数のデータを用いて、推定したい物理量を計算する際によく使われる。重み付き平均値 (the weighted mean) の計算式である。(2)式はその標準不確かさ (the standard uncertainty) である。データに補正すべきバイアスが残っていない場合はこの式でよいが、そうでない場合の外的不確かさ (the external uncertainty δ_{ext}) を後に考えるため、ここでは内的不確かさ (the internal uncertainty δ_{int}) と定義しておく。 δ_{int} が推奨値の不確かさとして妥当か否かを判定する手法として、以下のカイ二乗検定がよく知られている。

$$\chi^2 = \sum_{i=1}^n \frac{(A_i - \bar{A})^2}{\delta A_i^2} \quad (4)$$

カイ二乗の期待値は、データ数から自由度 1 を差し引いた (n-1) となる。カイ二乗の値が(n-1)より大きい場合は、補正すべきバイアスが残っていることが疑われることになる。(2)～(4)式を見比べると、内的不確かさ (δ_{int}) と外的不確かさ (δ_{ext}) はカイ二乗 χ^2 を用い、次式で関係づけられる。

$$\delta_{\text{ext}} = \sqrt{\frac{\chi^2}{n-1}} \times \delta_{\text{int}} \quad (5)$$

内的不確かさ (δ_{int}) と外的不確かさ (δ_{ext}) の大きな方を採用するという考え方は、一律に内的不確かさ (δ_{int}) を採用する手法に比較し、データ間のばらつきを考慮できるというメリットがある。一方、常に外的不確かさ (δ_{ext}) を採用してはどうかとの考えも生じるかもしれないが、この場合は、たまたまデータが近い値になると不確かさを過小評価しかねない。偶然2つの値が一致すると、不確かさがゼロというおかしな結果になる。どちらか大きい方を採用するという手法について、遠慮なくものをいう SG-41 の参加者からも特に反論はでなかった。

数式がたくさん出てきたので、このあたりで本手法を適用した演習問題を一つ、具体的な数値と図で示したい。データ間のばらつきが大きく残っている場合、推奨値の不確かさとして内的不確かさを用いるか、外的不確かさを用いるかによって大きな違いが生じる様子が、この演習問題で示されている。解答図の縦軸は、比較しやすいよう、ピーク値が1となるよう規格化している。この演習問題では簡単のため、2個の測定値が独立な場合を扱っている。同じ崩壊データをサンプル量の決定に採用しているなど、測定値に相関があり、かつ、その不確かさ全体への影響が大きい場合は、相関を考慮して不確かさを導出する必要があるので注意していただきたい。なお、相関の扱いについては小林捷平氏による「共分散は面白い」という入門的解説 [7] の他、素粒子物理学分野にもよい解説 [12] がある。

【演習問題】

採用すべきデータの絞り込みの結果、以下の 2 個の独立な測定値が得られた場合を考えよう。

$$A_1 = 10.0, \delta A_1 = 1.0$$

$$A_2 = 14.0, \delta A_2 = 1.0$$

この情報をもとに、推奨値とその不確かさを求め、結果を図で示せ。

【解答】

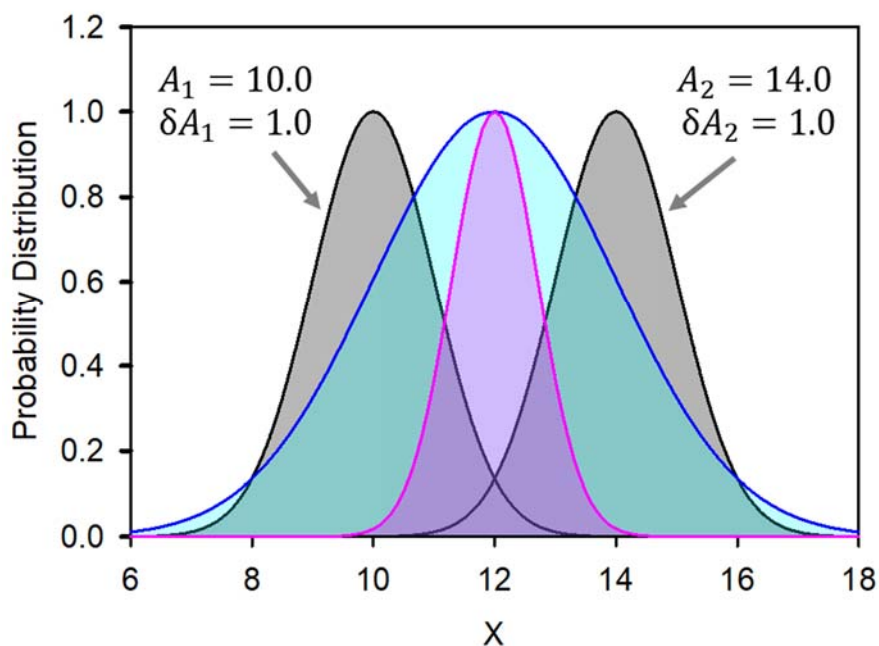
重み付き平均値 \bar{A} 、内的不確かさ (δ_{int})、外的不確かさ (δ_{ext}) は以下となる。

$$\bar{A} = 12.0, \delta_{\text{int}} = 0.7, \delta_{\text{ext}} = 2.0$$

$\delta_{\text{ext}} > \delta_{\text{int}}$ より、推奨値とその不確かさは、以下となる。

$$A = 12.0, \delta = 2.0$$

この結果を下図に示す。水色の分布が推奨値とその不確かさである。比較のため、不確かさとして内的不確かさ (δ_{int}) を用いた場合を紫色で示す。



内的不確かさ (δ_{int}) と外的不確かさ (δ_{ext}) を用いて不確かさの評価手法を提唱したのは、Raymond Birge 氏による 1932 年の論文 [8] が初めてであろう。この論文に次のような記載がある。「不確かさの予測に internal consistency を用いることができ、その予測が正しいか否か external consistency により知ることができる。Internal consistency と external consistency を比較し、external consistency の方が大きいならば、系統的な不確かさ要因があるはずであるから、この場合は external consistency に基づいて、不確かさを評価すべきだ。」と。この論文では不確かさの計算に、(3) 式の external uncertainty (δ_{ext}) ではなく、これに因子 $F(=0.6745)$ を乗じた式を用いている。この因子 F は、68.3% をカバーする今日の不確かさの定義に対し、当時は 50% をカバーする定義であったことを考慮すると定量的に理解できる。このように SG-41 で採用した導出式と細かな定義は若干異なるが、考え方の起源と考えられる。

同氏は、基礎物理定数や原子質量について、最良推定値 (the most probable value) のみならず、その推定誤差 (probable error) の導出に一早く取り組んだ研究者である。1929 年に発表した「Probable values of the general physical constants」というレビュー論文 [9] の中で、実験値のばらつきの問題を“a surprising lack of consistency”と認識し、今日に続くこの問題の難しさを早い段階で指摘している。

脱線するが、同氏はこの論文中に「誤差の誤差」あるいは「誤差の不確かさ」を論じており、不確かさは 2 桁の有効数字で表示することを推奨している。例えば、 126.944 ± 0.046 のように。現在、国際度量衡局 BIPM は、 1σ に相当する不確かさをカッコ内に示すことを推奨 [10] しており、これに従えば、 $126.944 (46)$ となる。

実験データのばらつきを不確かさの導出に反映させる方法として、現在、Birge 氏の手法を改定し、何種類もの手法が提唱されている。最近の物理標準分野での研究論文 [11] にある不確かさ導出手法に関する歯切れの悪い言い回しからは、まだまだ検討の余地が残っているように思われる。

素粒子物理学の分野でも、ばらつくデータの扱いは重要であり、大変参考となる解説 [12] がある。Birge 氏の手法が基本となり、その上で様々な議論が展開されている。一方、データが既存データに影響される効果 (Bandwagon effects と命名されている) を懸念した研究 [13] などもあり、なかなか一筋縄ではいかないようである。解説 [12] の序論も以下の文章で締めくくられている。「我々のデータ処理方法は通常は有効なのだが、不確かさを超えて外れることも起こり得る、いや必ず起こるものと認識しておくことは重要である」と。

5. 最終的な結果

最終的に、積分実験としては放射化法による複数の測定値を補正して得られた重み付き平均値が残り、微分実験としては TOF 法による複数の測定値から求めた重み付き平均

値が残った。今回のデータの絞り込みにより、 δ_{ext} は δ_{int} よりも小さくなった。文献 [1,14]でも引用したように、Ernest Lawrence氏が命名した Intellectual phase locking という思考停止に落ち込まないよう気を付けたつもりであるが、どうだったであろうか？後年の批判を待ちたい。なお余談だが、この Lawrence氏は、本稿で度々登場した Birge氏と深い絆がある。

さて、以下の表に結果をまとめる。ちなみに、微分実験を個別に見ると、最も不確かさが小さいのは、SG-41の報告書執筆時点で最新の ANNRIの測定値 [15]であり、その値は 707 (29) [b]である。この最も精度の高い最新の微分実験の精度に比較しても、全てのデータを集約して得られた精度は2倍以上高く、不確かさは 13 [b]まで絞り込まれた。国際協力による評価活動の大きな効果を示すものである。

SG-41による ^{241}Am の熱中性子捕獲断面積

	σ_0 [b]	δ_{int} [b]	δ_{ext} [b]
微分実験平均	720 (21)	21	17
積分実験平均	716 (16)	16	3
両手法の平均	717 (13)	13	2

この国際協力活動の前に、 ^{241}Am の熱中性子捕獲断面積を評価した研究論文がいくつかあるが、1つだけ紹介したい。それは中性子共鳴の教科書もある Eric Lynn氏による評価 [16]で、その値は 618 (20) [b]である。SG-41の推奨値に比べ、 5σ もの差異がある。Lynn氏の論文を読む限り慎重な議論が与えられている。評価に用いることのできた 1980年以前のデータの制約が差異の要因であったろうと推察される。データの制約には、測定の精度 (Precision) の低さだけでなく、影響の大きいバイアス要因に関する知見の不足もあったであろう。高精度化のためには、精度の高い測定データの蓄積とともに、補正すべきバイアス効果の知見の蓄積も大切である。

6. おわりに

ばらつきのあるデータにどのように向き合ったのか、核データ評価に関する国際協力として実施したサブグループ活動 (SG-41) を振り返り、試行錯誤の様子と若干の解説を思いつくままに書かせて頂いた。最後までこの乱文に目を通していただいた皆様には、何か一つでも参考となる処があれば幸いである。筆者の薄学により、考え違いや参照すべき先達の知見の抜けなど、多々あると思われる。ご叱正・ご教示を賜ることができれば幸甚である。

なお、本稿のきっかけは、大塚直彦編集長の依頼による。本活動への高い関心と不確かさ解析に関する有益な議論を頂いたことに感謝したい。

参考文献

- [1] H. Harada, JAEA-Conf **2013-002**, 47 (2013).
- [2] Nuclear Science, **NEA/NSC/R(2020)2** July 2020.
- [3] N.C. Barford, (酒井英行訳)「実験精度と誤差」丸善 (1997).
- [4] 吉澤康和, 核データニュース, **No. 8**, p.1 (1979).
- [5] 吉澤康和, 「新しい誤差論」共立出版 (1989).
- [6] S.F. Mughabghab,“Atlas of Neutron Resonances 6th Ed.”, Elsevier (2018).
- [7] 小林捷平, 核データニュース, **No. 100**, p.45 (2011).
- [8] R. T. Birge, Phys. Rev. **40**, 207 (1932).
- [9] R. T. Birge, Phys. Rev. Supplement **1**, 1 (1929).
- [10] Joint Committee for Guides in Metrology, JCGM 100:2008.
- [11] B. Toman et al., Metrologia **49**, 567 (2012).
- [12] S. Eidelman et al., Phys. Lett. B **592**, 1 (2004).
- [13] M. Jeng, Nucl. Instru. Meth. A **571**, 704 (2007).
- [14] Nuclear Science, **NEA/NSC/WPEC/DOC(2014)446** February 2014.
- [15] K. Terada et al., J. Nucl. Sci. Technol., **55**, 1198 (2018).
- [16] J.E. Lynn et al., Prog. Nucl. Eng. **5**, 255 (1980).