平成 25 年度　分析化学学生実験

データ解析

はじめに

近代化学を飛躍的に進歩させた技術の一つに X 線回折を利用した分子構造解析がある．それまでマクロな視点でしか推定し得なかった分子の構造を初めて目に見える描像として浮き上がらせたのである．しかし、X 線回折技術はオングストローム単位の分子を目に見えるまで拡大する魔法ではない．得られた回折像を眺めても分子の形はちっともハッキリしない．これを解析し、対象性やらセル定数やらを決定していき、最終的に得られた回折像を再現するような分子構造を推定するのだ．

今日、化学の分野においてデータを加工して欲しい情報を抽出する作業は常について回る．加工前のデータを「生（なま）データ」と言ったりもする──英語でも*raw data* と言うので、おそらく世界共通の概念なのだろう．これから始まる分析化学学生実験においても、大量の測定値をいくつかのパラメータに集約する局面が出てくる．こういったデータ処理には慣れていないであろうから、分析化学学生実験の最初にまずデータ解析の単元を設け、統計学の基礎と、パーソナルコンピュータを用いたデータの取り扱いについて触れてもらう．ともすれば Excel の使い方に慣れるのに終始してしまうかもしれないが、ともかくここで自力でデータ解析できるようになってもらわなければ以降の実験が無意味なものになってしまうので、しっかりと習得して欲しい．

第一章：誤差

1-1. 正規分布

実験データには誤差がつきものである．多くの場合、測定値は“*真の*”値の前後に**正規分布**と呼ばれるある法則に則ってバラついているらしい．正規分布は関数として与えられているが、これを導き出すのに多くの近似を用いているし、もともとの理論的裏付けもちょっと弱いところがある．しかしその後の数学的取り扱いが比較的容易であることと、何より多くの場合に測定値のバラつきをうまく再現するため、実験誤差を取り扱う上で必要不可欠な理論・関数となっている．

まずは見た目から入ろう．図 1 に、$x = 0$ を中心とした正規分布を示した．すその広がった釣鐘状をしていることが分かる．正規分布関数は 2 つのパラメータ（関数の形を決める変数）mとs^2 からなり、$N(m, s^2)$と表すのが慣例となっている．N は Normal distribution の頭文字である．mおよび s^2 の意味は後で明らかになる．関数 $N(m, s^2)$の実態はこうである．

$$N(m, s^2) = \frac{1}{\sqrt{2p\, s^2}} \exp\left\{-\frac{(x-m)^2}{2s^2}\right\}$$

exp(a)は e^a を表す．指数が複雑になると小さくてわかりにくいので、このように書くことになっている．パーソナルコンピュータを用いた解析においても同様の使い方をするので、この表記法にも慣れて欲しい．なぜこのような複雑な式となっているかは後で分かる．まずは 2 つのパラメータからなる釣鐘型の関数であるということを覚えておいて欲しい．

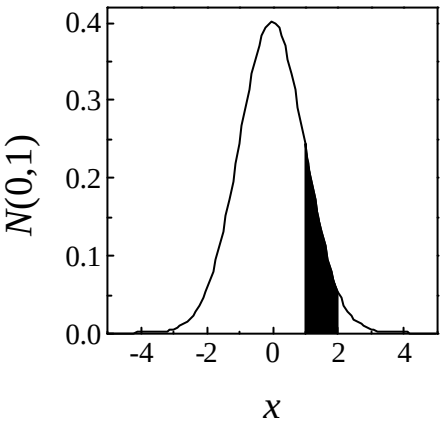

図 1標準正規分布曲線．黒塗りの部分は、$1<x<2$ で、面積は 0.136.

縦軸は頻度を表すのだが、その絶対値にはあまり意味がない．ある x の範囲で挟まれた区間の面積が、x がその範囲内に出現する確率を示す．例えば、

1. x = 1 のとき 0.242 である．これは、x が 1 ± 0.005 である確率が 0.242×0.01 = 0.24%であることを示している．（dx = 1.005–0.995 = 0.01、ただし dx は十分小さいことが前提である）
2. 黒く塗った部分（$1 < x < 2$）の面積は 0.136 である．これは、$1 < x < 2$ である確率が約 14%であることを示している．

面積で確率を示すような関数を**確率密度関数**と呼ぶ．正規分布は確率密度関数の中で最も代表的な一つである．

面積が確率を示すということは、x を全範囲（$-\infty < x < +\infty$）にわたって積分すると 1 になるということだ．これは確率密度関数の最も大切な性質に挙げられる．正規分布でこのことが成り立っているのだろうか、確かめてみよう

ワンポイント

x が 1 ピッタリである確率が 0.242 というわけではない．例えばもし下 2 桁まで表示できるデジタル機器で値が 1.00 と出たら、実際は $0.995<x<1.005$ ということである．

■宿題 1：以下の積分を解け

$$\int_{-\infty}^{+\infty} N(0,s^{2})\mathrm{d}x$$

データ解析の時間に指名し、解答を板書してもらう．当てられても良いように、以下のヒントを参考にして、各自で式変形をしておくこと

1. $x = 0$ で対称なので、そこで分けて積分する．すなわち、

$$\int_{-\infty}^{+\infty} N(0,s^{2})\mathrm{d}x = 2\int_{0}^{+\infty} \frac{1}{\sqrt{2\pi s^{2}}}\exp\left(-\frac{x^{2}}{2s^{2}}\right)\mathrm{d}x$$

2. 以下の定積分の公式を用いる．

$$\int_{0}^{+\infty} \exp(-ax^{2})\mathrm{d}x = \frac{1}{2}\sqrt{\frac{\pi}{a}}$$

ワンポイント

s^{2} は s^{2} であって、何かの値の2乗ではない．だから式変形の過程で平方根を取って s に書き直す必要はない．

ワンポイント

積分範囲が $-\infty<x<+\infty$ であるので、m が有限の値であればそこで分けて積分することで同じ結果が得られる．

無事に積分値が 1 となったであろうか．m が有限の値であれば、一般形 $N(m, s^{2})$ について成り立つ．

さてある測定値が正規分布 $N(m, s^{2})$ に則って分布した場合、その期待値 $\langle x\rangle$ はいくらになるであろうか．期待値は全範囲における値と事象との積の和 $\langle x\rangle = \int_{-\infty}^{+\infty} xN(m,s^{2})\mathrm{d}x$ で表される．

ワンポイント

期待値は、ある値とその事象の起こる確率の積の総和である．例えばさいころの目の期待値は 1/6+2/6+...+6/6 = 3.5 である．

■宿題 2：以下の積分を解き、正規分布の $\langle x\rangle$ を導け

$$\int_{-\infty}^{+\infty} xN(m,s^{2})\mathrm{d}x$$

ただし、以下のヒントを参考にせよ．

1. 変数変換せよ．例えば、$u = x - m$ として、

$$\langle x\rangle = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi s^{2}}}(u+m)\exp\left(-\frac{u^{2}}{2s^{2}}\right)\mathrm{d}u$$

と変換する．このとき、$\mathrm{d}x = \mathrm{d}u$ である．
積分範囲は $-\infty < u < +\infty$ で変わらない．

2. 奇関数の $-\infty$ から $+\infty$ までの積分は 0 となる．すなわち、

$$\int_{-\infty}^{+\infty} u\exp\left(-\frac{u^{2}}{2s^{2}}\right)\mathrm{d}u = 0$$

3. 宿題 1 で導いた関係式も使える．すなわち、

$$\frac{1}{\sqrt{2\pi s^{2}}}\int_{-\infty}^{+\infty} \exp(-\frac{u^{2}}{2s^{2}})\mathrm{d}u = 1$$

結局、$\langle x \rangle = m$が得られるはずである．このmを**母平均**と呼ぶ．正規分布$N(m, s^2)$の最初のパラメータmは母平均と等しい値であった．mと$\langle x \rangle$は全く別物であり、mはあくまで分布の特徴を示すパラメータ、$\langle x \rangle$は実測値の期待値である．我々は測定結果を解析することによって$\langle x \rangle$を知ることができるが、これより得られたmは推定値でしかない．

ワンポイント

母集団と**標本**は似て非なるものであるという考え方をここでは徹底しよう．

つぎに、その測定値はどれぐらいバラついているであろうか．バラつきの指標として、**分散**と呼ばれるs^2という値を導入する．s^2は、測定値 x の期待値$\langle x \rangle$との差の二乗の期待値として定義される．

$$s^2 = \langle (x - \langle x \rangle)^2 \rangle$$

宿題 2 によると、$\langle x \rangle = m$であるから、分散s^2は

$$s^2 = \int_{-\infty}^{+\infty} (x-m)^2 N(m, s^2) \mathrm{d}x$$

で表される．

ワンポイント

当たり前だが「差の期待値」は 0 である．だから二乗した．絶対値でも良かったのだが、絶対値はとかく数学的に取り扱いにくい．

■宿題 3：以下の積分を解き、正規分布の分散を導け

$$\int_{-\infty}^{+\infty} (x-m)^2 N(m, s^2) \mathrm{d}x$$

ただし、以下のヒントを参考にせよ．

1. $u = x - m$と変数変換せよ．このとき、$\mathrm{d}x = \mathrm{d}u$ である．
2. $f = -s^2 \exp\left(\dfrac{-u^2}{2s^2}\right)$，　$g = u$ とおき、

   積分公式$\int_a^b f' g \mathrm{d}x = [fg]_a^b - \int_a^b fg' \mathrm{d}x$ を使う
3. 宿題1で導いた関係式、および宿題2のヒント3の関係式が使える．

ワンポイント

繰り返すが、期待値とはある値（ここでは$(x-m)^2$）とその確率（ここでは $N(m,s^2)\mathrm{d}x$）の積の総和である．

勘の良い人は読めていたであろうが、$s^2 = s^2$が得られるはずである．このs^2を**母分散**と呼ぶ．正規分布 $N(m, s^2)$の 2 番目のパラメータは正規分布の母分散であった．これはもちろん偶然ではなく、式の中にうまい具合に母平均と母分散を埋め込んだためである．定義式が少し不自然な形をしていた理由がわかったであろう

分散s^2も、正規分布のパラメータs^2とは由来が異なるものであることをよく理解されたい．我々は複数回の測定（サンプリング）を行うことによって s^2を計算し、これをもってもともとの分布の母分散s^2を推定するのである．

また、分散の平方根を**標準偏差**といい$\sqrt{s^2}$で表す．そうすることによって測定値 x と次元が同じになり、値が意味を持つようになる．標本（測定値）の標準偏差は慣例的にs_xと書くが、あくまでもともとの定義は分散の平方根であることを忘れないようにしたい．

1-2. データの確度

これまで、値の分布のしかたについて学んだ．もちろん、全てのデータが正規分布するとは限らないが、それでも正規分布は最も重要な確率分布関数である．ここでは、誤差が正規分布するデータの取り扱いについて、より定量的・実践的な方法について述べる．

正規分布は重要な確率密度関数であるが、実際の値を計算するのは容易なことではない．普通の電卓ではお手上げである．そこで、誤差を取り上げている教科書の多くは $N(0, 1)$ の値の一覧が掲載されている．$N(0, 1)$ がわかれば、一般形 $N(m, s^2)$ も簡単に計算できるのである．

■宿題 7：$N(0, 1)$ から $N(m, s^2)$ を導け

$$f°(x)=\frac{1}{\sqrt{2p}}\exp\left(\frac{-x^2}{2}\right),\quad f(x)=\frac{1}{\sqrt{2p\,s^2}}\exp\left\{\frac{-(x-m)^2}{2s^2}\right\}$$

とおき、$f°(x)$ を用いて $f(x)$ を表せ．

ヒントは…$f(x) = \alpha f°(X)$ となるような X を探そう

$f(x)=\frac{1}{\sqrt{s^2}}f°\left(\frac{x-m}{\sqrt{s^2}}\right)$ が得られたであろうか．すなわち、$N(0, 1)$ を移動し、縦横に拡張・圧縮すれば、任意の一般形 $N(m, s^2)$ に自由に変換できるのである．この $N(0, 1)$ は特に**標準正規分布**と呼ばれている．

ワンポイント

冒頭で正規分布は数学的取り扱いが容易だと書いたが、首を傾げたくもなってしまう．他にもっと便利で、測定結果を良く再現して、数学的取り扱いも容易で、かつ計算も簡単な関数はないものか．

確率密度関数で興味があるのは、その絶対値（確率密度）よりも、確率そのものである積分値 $\int_a^b N(m,s^2)\mathrm{d}x$ である．しかし残念ながら、この定積分は解析的に解くことができない．そこで、$N(0, 1)$ と同様、$-\infty$ から a までの積分値 $\Phi°(a)=\int_{-\infty}^{a} N(0,1)\mathrm{d}x$ も載っている場合がある．この関数を用いて、任意の区間の積分値を得ることができる

$$\int_a^b N(0,1)\mathrm{d}x=\int_{-\infty}^{b} N(0,1)\mathrm{d}x-\int_{-\infty}^{a} N(0,1)\mathrm{d}x$$
$$=\Phi°(b)-\Phi°(a)$$

確率密度関数の積分の関数を、誤差関数または累積分布関数と呼ぶ．図 2 に標準正規分布の誤差関数を描いた．0 から始まり 1 に漸近している．図によると、$\Phi°(1) = 0.841$、$\Phi°(2) = 0.977$ であるので、$1 < x < 2$ である確率は $\Phi°(2) - \Phi°(1) = 0.136$（14%）である．図 1 の黒く塗った面積はこのようにして計算した．

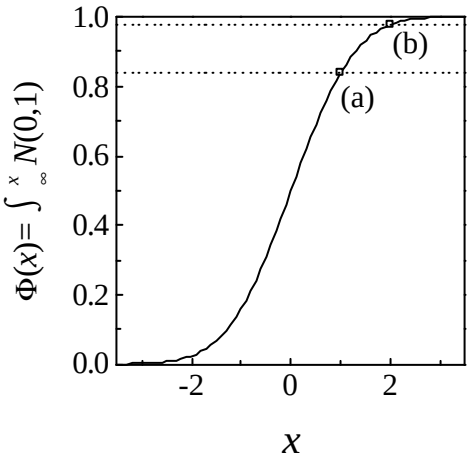

図 2 累積正規分布曲線．点 (a)および(b)の x 座標は各々1 および 2 である．

標準正規分布の誤差関数も、標準正規分布と同様、任意の正規分布に変換することができる.

ワンポイント
0 から a の積分値の 2 倍、または $-a$ から a までの積分値が掲載されている場合もあるので、確認して使うこと.

■宿題 8：$\int_{-\infty}^{a} N(0,1)\mathrm{d}x$ から $\int_{-\infty}^{a} N(m,s^2)\mathrm{d}x$ を導け

$$\Phi°(a)=\int_{-\infty}^{a} N(0,1)\mathrm{d}x$$

$$\Phi(a)=\int_{-\infty}^{a} N(m,s^2)\mathrm{d}x$$

とおき、$F°(a)$を用いて$F(a)$を表す.

ヒント：$u=\dfrac{x-m}{\sqrt{s^2}}$ とおき、変数変換を行う.

ちなみに、正規分布の誤差関数は解析的に解けないと書いたが、以下の関数でかなり良く近似できる.

$$\Phi(x)\approx\frac{1}{2}\pm\frac{1}{2}\sqrt{1-\exp\left(\frac{-2x^2}{p}\right)}$$

ワンポイント
プログラムを書いたりするときには必要であるので、そういう関数があることぐらいは覚えておくと良いかもしれない. 一般の表計算ソフトには組み込み関数として提供されている.

標準偏差を定義したときに、x と比較できて便利であると書いた. いま、標準正規分布について考えてみよう $\Phi°(s_x) = \Phi°(1) = 0.841$、$\Phi°(-s_x) = \Phi°(-1) = 0.159$ であるから、$-s_x < x < s_x$ である確率は $\Phi(1) - \Phi(-1) = 0.683$ である. すなわち、x が標準正規分布に則っている場合は、x の絶対値が標準偏差以内である確率は約 7 割であるということが言える. 一般形ではどうだろう.

■宿題 9：$\Phi(m\pm s)$ を $\Phi°(\pm1)$ を用いて表せ

宿題 8 を解けていれば. これに、$x = m \pm \sqrt{s^2}$ を代入するだけで簡単に導ける.

どんな $N(m, s^2)$ においても $\Phi(\pm s_x)$ は $\Phi°(\pm1)$ と等しいことが明らかとなった. すなわち、x が正規分布していれば、その期待値からの差が s_x^2 以内に収まる確率は 68.3% である、と言える. 標準偏差にはこのような意味があった.

さて、ある測定値の平均値が 3.1416、標準偏差 $s_x = 0.038$ であったとしよう. 先の考えに従うと、測定値の信頼性は、68% の確率で 3.1036～3.1796 の範囲である. どうやら、3.10 と 3.18 の間であるとは言えそうだが、それより先の、最後の 2 桁はてんでバラバラと思われる. 標準偏差も最初の桁以降は意味がないので四捨五入して $s_x = 0.04$ としか言えない. 結局、こういう場合は、3.14 ± 0.04 とまでしか書けない. 標準偏差と同じ桁までは有効であるが、それ以降はランダムであろうから、切り捨てる. これが**有効数字**の考え方だ. 結局、最初に計算した平均値 3.1416 のうち、有効数字は最

ワンポイント
標準偏差はその値そのものが誤差である. 誤差の最初の 0 でない桁以降はバッサリ切る.

初の 3 桁だけであった．また、今回はsで評価したが、より精度の高い議論を行いたいときは sの 3 倍を誤差範囲として用いる．その場合は $3s$ = 0.1 となり、結局有効数字は 2 桁だけの 3.1 ± 0.1 ということになる．データを真摯に眺めるとそういうことになる．その測定値が十分に精度が高いのか、実験の精度をより高めなければならないかはこれを見て判断する．

■宿題 10 2σと3σの信頼性はどれぐらいか

少しだけ視点を変えてみよう．図 3 に、この測定値の分布 $N(3.1416, 0.038)$を $3.1416 \pm a$ で積分したものを示した．縦軸はそのまま、x が$\pm$ a の範囲内にある確率である．a = sのとき 0.68 である．測定値がこのような分布をしているとき、a がsの内側にある確率はどんどん小さくなっていく．3.14 までぴったり合う確率 $\Phi(0.01)$は 20%にまで落ち込む．それ以下の桁なんて、合うことの方が珍しい．平均値なんて電卓叩くだけで 10 桁、12 桁まですぐに計算してくれるが、だからといってその値を闇雲に書き写せば良い訳ではない．

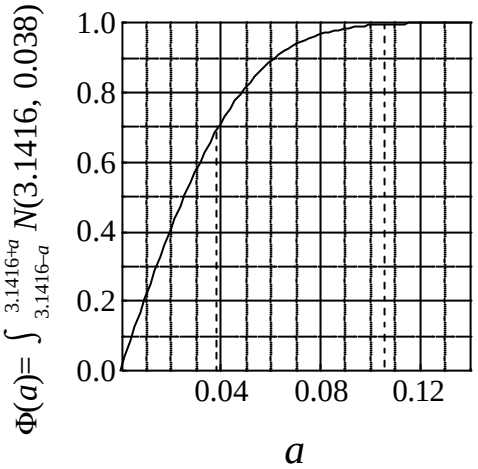

図 3 $N(3.1416, 0.038)$の累積正規分布曲線．縦線は、x = s (0.038)および x = $3s$ (0.114)

また、誤差を単に$\pm$ 0.04 と書いてあってもそれがsなのか $3s$なのか、それともその他の誤差評価関数なのか判断ができないため、明記する必要がある．例えば学術論文などでは、表中に 3.1 (0.1)などと書き、欄外に「The values in parentheses refer to three standard deviations」のように付け加えられる．誤差範囲を表すのに「±」なんて曖昧な記号は避ける傾向にある．

1-3. 誤差の伝播

繰り返し測定して、平均値とバラつき（誤差）を決定したら、それで終わりであるということは少ない．多くの場合、目的とする物理量を得るためには、より単純化した実験から得られた値に演算を施したり、あるいは複数の実験値と組み合わせたりする必要がある．そのとき、誤差はどのように伝播してゆくのだろうか？

いま、ある物理量 x を測定し、その値と誤差 $s(x)$を決定したとしよう．目的とする物理量 y が、x の四則演算から求まる場合、目的の値の誤差の評価は簡単である．例えば、$y = ax$（a は物理定数などで、誤差はないとする）であるとする場合、y の誤差 $s(y)$は、

$$s(y) = a\,s(x)$$

で求められる．$y = x + b$ の場合は、

$$s(y) = s(x)$$

で得られるであろう．このことは、グラフを描いてみると分かりやすい．図 4 を見てみよう．測定量である x の値が$\pm s(x)$ずれたときに、それに演算を施して得られる物理量 y の値はどこまでずれるか？すなわち、誤差がどのように伝播するのか？というのは、そのときの関数の傾きに依存すると言えそうである．

四則演算でなくても応用できるよう、これを拡張する．例えば、$y = \ln x$ を取り上げる．この関数は直線ではないが、ある x の値近傍では、その傾きは導関数で表されるであろう．すなわち、

$$s(y) = \frac{\partial}{\partial x}\ln x \cdot s(x)$$

である．分散はあくまで $s(x)^2$、$s(y)^2$ で定義される値であるから、結局、y の誤差 $s(y)$は、測定量 x の誤差 $s(x)$を用いて、

$$s(y)^2 = \left(\frac{\partial}{\partial x}\ln x\right)^2 \cdot s(x)^2$$

$$= \frac{s(x)^2}{x^2}$$

と表す事ができる．

ワンポイント

詳しくは後述するが、N 個の測定値 x_i から得られる母分散 s_x は、

$$s_x = \frac{1}{N-1}\sum (x_i - \bar{x})^2$$

である．ここではこの値を誤差 $s(y)$として採用する．

■宿題 11：以下の関数における y の誤差を表せ

- 関数 $y = \log x$
- 関数 $y = ax + b$
- 関数 $y = \ln x$ において、x=10、$s(x)$=0.5
- 地球の直径を誤差 1m で測定したときの赤道上の地球一周の距離の誤差はいくらか？

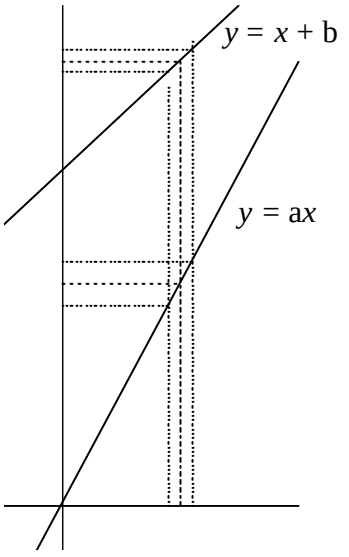

図 4 $N(3.1416,\ 0.038)$の累積正規分布曲線．縦線は、x = s (0.038)および x = 3s (0.106)

次に、複数の測定値から物性値を求めるときについて考えよう 例えば、2 つの測定値 x とy を加えて別の物性値 z を求める場合、z の誤差はどう考えれば良いか？単純に x と y の誤差を足し合わせればいい？では積の場合は？除算の場合は？

複数の測定値からなる場合も同様に、誤差を求めたい関数 z が測定値 x および y にどれだけ 敏感」であるかは、各々の変数による偏微分から見積もる.

$$s(z)^2 = \left(\frac{\partial z}{\partial x}\right)^2 \cdot s(x)^2 + \left(\frac{\partial z}{\partial y}\right)^2 \cdot s(y)^2$$

この式は誤差伝播則と呼ばれる．この式はもちろん、

$$s(z) = \left(\frac{\partial z}{\partial x}\right) \cdot s(x) + \left(\frac{\partial z}{\partial y}\right) \cdot s(y)$$

から出発し、

$$s(z)^2 = \left[\left(\frac{\partial z}{\partial x}\right) \cdot s(x) + \left(\frac{\partial z}{\partial y}\right) \cdot s(y)\right]^2$$

$$= \left(\frac{\partial z}{\partial x}\right)^2 s(x)^2 + 2\left(\frac{\partial z}{\partial x}\right)\left(\frac{\partial z}{\partial y}\right) s(x)s(y) + \left(\frac{\partial z}{\partial y}\right)^2 \cdot s(y)^2$$

詳しい説明は省略するが、このうち第 2 項は 0 となるり、よって、誤差伝播の式が導かれる．これより多い測定値を組み合わせた場合でも同様である.

■宿題 12 :以下の関数における z の誤差を表せ

- 関数 $z = ax + by$
- 関数 $z = axy$
- 関数 $z = ay/x$
- 関数 $z = a \exp(bx)$
- 関数 $z = a \ln(bx)$

第二章：最小二乗法を用いたデータフィッティング

二章は、大量のデータを統計処理するための理論を扱う. できるだけ正確なデータを得るために、複数回のサンプリングをして平均値を計算することぐらいはやったことがあるだろう. これもデータの統計処理のうちの一つである．この章ではまずその操作から入る．しかしそれは単なる相加平均とはやや意味合いの異なるものだ．これを足がかりとして、直線のフィッティング、および多変量解析について述べたい.

2-1. 平均値と分散

繰り返しになるが、測定には誤差がつきものである．例えば中和滴定で濃度を決定する時に、1 回だけで決定するなんてのはナンセンスだ．何かの拍子に突拍子もない値を示しているかもしれない．そもそも濃度決定なんてそれ自体が目的でない場合が多く、そこでコケると後の実験が全てパーになってしまうなんてことがザラだ．そこで同じ操作を何度か繰り返して、どれくらい正確な値が得られたかも議論する.

全体で N 回の測定を繰り返し、測定値が x_i ($i = 1, 2, ..., N$) であったとしよう．ここで、x の平均値 $\bar{x}$ はご存知のとおり、

$$\bar{x} = \sum_{i=1}^{N}$$

で表される．ここでちょっと発想を変えてみよう．ある期待値 a を仮定する．その a が最も実験値を良く再現するとき、その a こそが最も確からしい値だと言えよう．では「最も良く再現する」というのをどう定義するか？x_i との差の絶対値の総和 $\Sigma|x_i - a|$ が最も小さくなる a、でも良いが、ここでは後々の数学的取り扱いの容易さも鑑みて、x_i との差の二乗の総和が最も小さくなる a と定義することにする．この差の二乗の和を「残差二乗和」と呼び、U で表す．同じような方法で最も尤もらしい定数を見つける方法を最尤法と呼び、その評価関数に残差二乗和を用いる場合を最小二乗法と呼ぶ.

ワンポイント

絶対値というのはとかく数学的取り扱いがしにくい.

考察のヒント

評価関数に残差二乗和を用いるのが妥当かどうかは議論のしどころである. U は分散と関係する値であるが、分散も数学的背景に乏しかった.

さてこれから最小二乗法の手順を示そう．まず U の定義は、

$$U = \sum_{i=1}^{N} (x_i - a)^2$$

である．ここで、あっと驚く新発想、a が最も尤もらしい値の時に U は最小値かつ極小値をとる――すなわち、a はついて偏微分して 0 となるときの a が今回の最小二乗法の解となる. Σの中身が放物線であるので、この命題は正しそうである.

では早速微分してみよう.

$$\frac{\partial U}{\partial a} = \sum_{i=1}^{N} 2\left(\boxed{}\right)\frac{\partial}{\partial a}\left(\boxed{}\right)$$
$$= -2\sum_{i=1}^{N}\left(\boxed{}\right)$$
$$= -2\left(\sum_{i=1}^{N}\boxed{} - Na\right)$$
$$= 0$$

ヒント

$\frac{\mathrm{d}}{\mathrm{d}x} f(g(x))$ の微分の公式、よもや忘れてはいまい．$u=g(x)$などと変数変換して解くと、$g'f'(g(x))$となる.

また、$\frac{\partial}{\partial x}$をΣの中に入れて良いかどうか、分からなくなったら展開して考えること.

これを解くと、

$$\sum_{i=1}^{N} x_i = Na$$

$$\therefore a = \frac{\sum_{i=1}^{N} x_i}{N}$$

となり、結局は平均値と同じ式が得られる．なーんだ、という声が聞こえてきそうだが、今回はちゃんと理論的裏付けの下に「最も確からしい値」として a を得たのだ．それがたまさか、いつも使っていた「相加平均」と同じになっただけである.

測定値が正規分布をしているとしたら、その分散と同じぐらい x_i もバラついているであろう. x_i のバラつきからもとの分布の分散を求めよう. まず分散 s^2 の定義は、

$$s^2 = \langle (x - \langle x \rangle)^2 \rangle$$

であった. N 回の測定値において$\sqrt{s^2}$ に相当する量を計算すればそれが母分散のsであるだろう．ただし、N で割る代わりに$(N - 1)$で割る．これは、N 個のデータから1 個の変数αを求めたので、真に独立な値は$(N - 1)$個であるからである.

$$s^2 = \frac{\sum \left(x_i - \langle x \rangle\right)^2}{N-1}$$

分母についてはもっともらしいことを書いたが、実際には母集団の分散（母分散）から得られる測定値の分散の期待値を得る過程において$(N - 1)$の項が入ってくるのである．これを紙面で誘導するのは難しいし、数学遊びになってしまうので割愛する．興味がある人はやってみて欲しい.

ワンポイント

$(N-1)$で割る統計学的意味合いにおいてこの説明で当たらずとも遠からずといったところなので、そう理解しておいて構わないと思う.

それよりも、得られた a および s^2 ははあくまで測定値から推定した値であって、“真の”母集団のmおよびs^2 とは異なることに注意しよう. 今回得たのはあくまで母集団からの有限の標本から得られた「最も尤もらしい値」であって、母集団を全てサンプリングしたわけではない．もちろん、Nが大きくなればなるほど、標本は母集団に一致していく.

2-2. 原点を通る直線のフィッティング

次に、$y = ax$ の係数を求める場合について考える．今回の実験で言うと、濃度を変えて吸光度を測定したデータから吸光係数とその誤差を決定する際に用いるので、少し真面目に考えてみよう．このとき、データyは吸光度（実験では A を用いて表す）、要素 x は濃度（C で表す）、濃度あたりの吸光度（モル吸光係数eと光路長lの積である）が、係数 a に相当する．

ワンポイント
前回は、測定値だけの羅列であったから、x_i が i 番目の測定値、今回は x_i は i 番目の要素であり、測定値は y_i である．

まずは、最小二乗法を用いて、モル吸光係数eを決定する（これは関数の係数 a に相当する）．U の定義は、

$$U = \sum_{i=1}^{N} (y_i - ax_i)^2$$

であった．これを***a* について**微分すると、

$$\frac{\partial U}{\partial a} = \boxed{}$$

である．この式を用いて$\frac{\partial U}{\partial a} = 0$ を解くと、

$$a = \boxed{}$$

が得られる．また、標準偏差は、自由度（フィッティングすべき変数）が 1 つなので、$x = a$ のときと同じように

$$s^2 = \frac{\sum_i (y_i - ax_i)^2}{N-1}$$

で与えられる．

ワンポイント
a は N 個（データ数）の y_i の関数である、として考える．変数と定数が入れ替わることに注意しないと、何をやっているのか分からなくなるぞ．

さてここで、誤差がy_i にしかないとすると、伝播式により、a の誤差は

$$s(a)^2 = \sum \left(\frac{\partial a}{\partial y_i} \right)^2 s^2$$

となる．ここで、

$$\frac{\partial a}{\partial y_i} = \boxed{}$$

ゆえに、

$$s(a)^2 = \boxed{}$$

が得られる．

2-3. 一般的な直線のフィッティング

お次は最小二乗法を用いて N 個の平面上のデータ点 (x_i, y_i) に直線 $y = ax + b$ を当てはめる方法である．もしかしたらこれまでに、Σのやたら入った複雑な式を使って傾きと切片を求めたことがあるかもしれない．あれは実は最小二乗法だったのだ！今回のセクションはあの式を導き出すことが目的であると思ってよい．

ワンポイント

最小二乗法を使うと、グラフに定規を当てて傾きと切片を求めるよりも確実な値が得られる．

今回は、誤差は y 方向にのみあるとし、x 方向の値は十分に確からしいとしよう．まずは U の定義からだ．i 番目のデータ y_i について、前節の a の代わりに $ax_i + b$ を用いて、

ワンポイント

x、y 双方に誤差がある場合は、各々の値の確からしさをある程度正確に見積もっておかなければならない．

$$U = \sum_{i=1}^{N} \left(\boxed{} \right)^2$$

U が最小になるような a および b を求める．そこで、a および b について偏微分し、各々を0とおく．

ヒント

$u=y_i-(ax_i+b)$ とおいて微分すると良い．

$$\frac{\partial}{\partial a} \sum_{i=1}^{N} \{y_i - (ax_i + b)\}^2 = 0$$

$$\frac{\partial}{\partial b} \sum_{i=1}^{N} \{y_i - (ax_i + b)\}^2 = 0$$

先の場合と同様、a と b の偏微分であることに注意しよう．x や y で微分してはいけない．Σの中では x と y が変数であるので、これらはΣの外に出してはいけない．逆にΣの中で a と b は定数であるのでΣの外に出すことができる．

ヒント

わからなくなったら、Σを展開して地道に計算してみること．

$$\frac{\partial U}{\partial a} = -2 \sum_{i=1}^{N} \left(\boxed{} \right)$$

$$\frac{\partial U}{\partial b} = -2 \sum_{i=1}^{N} \left(\boxed{} \right)$$

$\frac{\partial U}{\partial a} = 0$、$\frac{\partial U}{\partial b} = 0$ より、a、b について整理すると、

$$\left(\sum_{i=1}^{N} \boxed{} \right) a + \left(\sum_{i=1}^{N} \boxed{} \right) b = \left(\sum_{i=1}^{N} \boxed{} \right)$$

$$\left(\sum_{i=1}^{N} \boxed{} \right) a + Nb = \left(\sum_{i=1}^{N} \boxed{} \right)$$

という連立二元一次方程式が得られる．この括弧内は全て実験値から計算できる値であり、既知の値である．そこで a と b について方程式を解くと、

$a =$

$b =$

が得られる．また、このときの分散s^2は

$$s^2 = \frac{\sum\{y_i - (ax_i + b)\}^2}{N-2}$$

である．これも導出は困難なので省略する．変数が2つになったため、分母が$N-2$となったことに注意せよ．これも標本分散から母分散を得る過程で顕わになるのだが、どのみち仮定が入るしNが十分に大きい場合はその違いは論ずるに足りないので、詳しくは成書に譲る．

ワンポイント

aとbは直線の見かけを決定するパラメータである．最小二乗法とはパラメータフィッティングである．

なお、傾きと切片の誤差 $s(a)^2$ および $s(b)^2$ は下記のとおりである．これも紙面での導出は困難だ．パラメータa および b に関する誤差伝播の式を適用するとことがミソである．

$$s(a)^2 = s^2 \frac{N}{N\sum x_i^2 - \left(\sum x_i\right)^2},\quad s(b)^2 = s^2 \frac{\sum x_i^2}{N\sum x_i^2 - \left(\sum x_i\right)^2}$$

もう少し簡単な直線の式について、宿題にするので考えてみよう$y = ax$、すなわち、必ず原点を通る直線のとき、a の値はどのように表されるか？まずはUを定義し、微分して=0となるxを探す、という、通りに解いていけば問題ないと思う

$a =$

なお、このとき、標準偏差は下記のようになる．

$$s(a)^2 = \frac{s^2}{\sum x_i^2}$$

参考図書

- 「データ解析―アナログとディジタル［改訂版］」、粟屋 隆 著、学会出版センター
- 「納得する化学数学」、佐藤 博保 著、講談社
- 「化学者のための数学十講」、大岩 正芳 著、化学同人
- 「計測における誤差解析入門」、John R Taylor 著、林 茂雄・馬場 凉 訳、東京化学同人

第三章：計算機を用いたデータ処理

二章まで読み進めてきた諸君は、その計算量の多さに慄いているのではないだろうか．データのフィッティングとしては 2 番目に簡単な直線の最小二乗法でさえ、全データの二乗の和を計算して、x とy の積の和を計算して…など、とても人の手に負えるものではない．それでも四則演算で済むうちは手元の計算機で計算できるのでまだ良いが、指数、対数が入ってくるとお手上げである．ここでは、Excel を用いたデータ解析について述べ、最終的には直線に限らずどんな形の関数にも最小二乗フィッティングできるようになることを目的としている．これはデータ解析に非常に強力な手段であるから、この先必ず使う機会が出てくる．もしその時に方法を覚えていなくても、このテキストを読み返して思い出してくれると嬉しい．同時に、グラフの描き方についても指南する．グラフは誰か（自分も含めて）に対してデータを視覚的に訴えるための方法だから、常に見やすいグラフを作成することを心がけたい．

3-1. 原点を通る直線のフィッティング

原点を通る直線に乗ることがわかっている 15 組の(x, y)のデータを与える．まずはこのデータをグラフにプロットする．次に、このデータに

$$y = ax$$

をフィッティングする．データ最も良くを再現するような a の値を誤差も含めて決定せよ. 理論線をグラフ中に書き込み、得られた a の値の信頼度について記せ.

3-2.直線のフィッティング

直線に乗ることがわかっている 15 組の(x, y)のデータを与える．このデータに

$$y = ax + b$$

をフィッティングする. a、および b の値を誤差も含めて決定せよ．グラフには理論線も書き込め．さらに、その切片の大きさ比 b/a を求め、その誤差も決定せよ.

3-3. 対数のフィッティング

次のデータは、関数 $y = ae^{bx}$ に則ることが分かっているとする．重要なことであるが、最小二乗フィッティングとはどのような関数に当てはめるかを探すための手法ではなく、関数形は理論的にわかっていて、その**パラメータを決定するための手法**である.

まずは実測値をプロットする．理論値をフィッティングして、パラメータa と b を得たいわけであるが、まだ我々は直線のフィッティングしか知らない．そこで、この関数を無理やり直線に変換する.

$y = ae^{bx}$
$\ln(y) = \ln a + bx$

x に対して直線の式を得ることができた．これを用いて最小 2 乗フィッティングを行い、y および $\ln y$ グラフに理論曲線を描け．

3-4. 非線形のフィッティング

これまでは、直線のフィッティングを扱ってきた．しかし実測値が直線でない、あるいは線形に変換できない例はいくつもある．今回はそのような例を取り扱う．これができるようになると、データ解析においてかなり強力な武器となるのでぜひマスターして欲しい．

今回は、関数形に $y = 0.5\exp(-ax) + 0.5\exp(-bx)$ を取り上げる．1 から始まる 2 成分の減衰曲線である．

まず、適当な値を推測する．この値を初期値として、残差二乗和 U が最小となる a と b を探す．この値は、似たような他の系を参考にしても良いし、グラフから概算しても良い．今回は、半減期が約 4 であるから、$(\ln 2)/4 = 0.17$ を初期値としよう．そこから、Excel の「ソルバー」機能を用いて、実験値を最も良く再現するような a および b を決定せよ．

ワンポイント

ここで用いる a と b は単に初期値であるから、理想的にはどんな値でも良い．しかし今から最小である U を“探す”という作業をするので、より確実な値を使うに越したことはない．

今回のデータは全て捏造されたデータである．あるパラメータを仮定し、それに適当なバラつきを与えた．用いたパラメータは以下の通りである．誤差によってどれだけもとの値とずれているであろうか．

セクション	関数形	a	b
3-1	$y = ax$	5	
3-2	$y = ax + b$	3	10
3-3	$y = be^{ax}$	3	0.2
3-4	$y = 0.5e^{-ax} + 0.5e^{-bx}$	0.25	0.05