間違いから学ぶ

実践統計解析【第2回】

正規分布を極める

川瀬　雅也[1]*・松田　史生[2]

本連載1回目では「正規分布」について説明しなかったので，今回は正規分布について解説を行う．なぜ，正規分布が統計処理を行う上で重要なのかを，まず，考えてみたい．本来は確率や確率分布の話を行わなければならないが，これらの話題は皆さんが持っている統計学の教科書などを見ていただきたい[1)]．

正規分布って何？

AさんとB君が再びX教授のもとを訪ねてきた．正規分布の話をする約束なのだ．

Aさん：こんにちは，教授．
X教授:いらっしゃい, 待っていたよ．何かデータを持ってきたかな．
B君：はい．僕たちが練習で測定した酵素活性の数値を持ってきました．
Aさん：私の練習に付き合ってもらったんですけど．

表1．AさんとB君の測定結果

Aさん				B君			
15.8	15.1	15.3	16.5	15.7	16.1	15.6	16.0
15.2	15.9	16.5	16.7	16.1	15.9	16.2	16.1
15.0	14.9			15.8	16.2		

X教授：さすがに，B君の方がバラつきが少ないね．では正規分布に従うのかどうか確認してみよう．
Rにデータを入れよう．

```
> A <- c(15.8,15.1,15.3,16.5,15.2,15.9,16.5,16.7,15.0,14.9)
> B <- c(15.7,16.1,15.6,16.0,16.1,15.9,16.2,16.1,15.8,16.2)
```

箱ひげ図を書いてみよう．

```
>boxplot(A,B)
```

X教授：いきなり質問なんだけど，このデータからまずは何を知りたいのかな？それからなんでデータがばらつくんだったっけ？
Aさん：もちろん酵素活性値を測定したいです．
B君：それから実験ごとのランダムな微小な誤差のせいでばらつきが起きると前回勉強しました．
X教授：じゃあ，仮に誤差がまったくない実験ができたとしようか．すると得られた酵素活性値は何度実験しても同じ値になるはずだよね．これが今回測定したい酵素活性値だ，というのは想像できるよね．
Aさん：でも，先輩でもそんな実験はできません．
X教授：もちろん誤差はなくせない．じゃあランダムな誤差をふくむ活性値をできるだけ多く，20回とかケチくさいこと言わずに1万回とか，さらには無限回測定したとしよう．その活性値データでヒストグラムを書くとどんな形になるとおもう？
B君：確か，それが正規分布になるんじゃないんでしたっけ？
X教授：その通り．B君，やるじゃないか．19世紀の数学者カール・フリードリッヒ・ガウス（独）は三角測量の誤差の研究で，誤差の分布が正規分布になることを発見した．その後，多くの自然科学の現象でも，同様の事実が見つかってきたんだ．
B君：要するに，僕たちが使うデータの統計処理は正規分布を前提にしてみようということですね．
X教授：かなり荒っぽいが，そう考えてもらってもいいだろう．ただし，生態なんかの分野では，正規分布にならないケースも多いので，あくまでも生物化学の範囲ということで考えてほしい．もう1点，母集団と標本の概念が大事なんだ．母集団とは，研究対象全体を

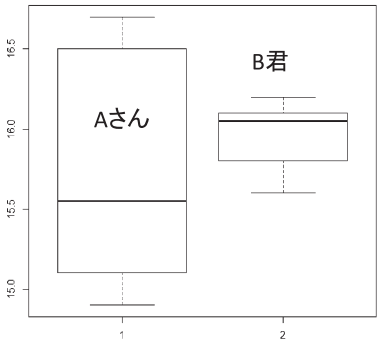

図1．AさんとB君の測定結果の箱ひげ図

*著者紹介　[1]長浜バイオ大学（教授）　E-mail: m_kawase@nagahama-i-bio.ac.jp
[2]大阪大学大学院情報科学研究科（准教授）

指す．たとえば，日本人男性の体重の平均値を調べるときは，実在する日本人の男性全員が母集団となる．酵素活性測定の場合は，無限回の実験から得た酵素活性値を母集団と考える．この場合母集団は仮想的なもので，実在しないんだ．

X教授：それからもう一つ重要なのは，普通の実験では母集団の平均値の計測を目指しているということだ．そこで，母集団の平均値のことを母平均と呼ぶ．生物工学の実験データでいえば，正しく実験が行われたことを前提とすれば「母平均」が「真の値」に当たると考えてよい．

Aさん：そっかぁ．母集団って仮想的な場合もあるんですね．統計の講義ではそこのところで混乱しちゃって，得られたデータ全部を母集団だと思っていました．

X教授：表1のような実験データのことを，母集団から取り出された標本と考えるんだ．母集団から偏りなく，母集団の性質を欠かさないように標本を取り出すことを無作為抽出という．10反復の実験は，無限個の測定値の母集団から10個を無作為抽出した，というふうに考える．

Aさん：ということは，正規分布の母集団から無作為抽出した標本だとしたら，表1の酵素活性値も正規分布にならないといけないはず，ということですね．

正規分布の基本式などは，皆さんの持っている統計学の教科書を見てもらうことにして，データが正規分布に従っていると考えていいかどうかの確認法を説明する．

正規性の確認

X教授：データが正規分布に従っているかを調べるにはQ-Qプロットという方法を使うんだ．これは，データの度数分布と，正規分布とした場合の度数分布とを比較して，両者がどのくらい似ているのかを見る方法だ．まず，Aさんから

```
> qqnorm(A)
> qqline(A)
```

Bくんのデータでも同じような処理をする．

この図は縦軸が測定データで横軸が理論値だが，直線にきれいにのると正規性が高いと判断できる．二人ともずれはあるが，おまけで何とか正規分布と見ていいという程度かな．

Aさん：おまけですか？

X教授：統計的に言うとShapiro-Wilk normality testを行ってみないといけないが，多分，計算すると二人とも正規性なしという結果になる．しかし，生物工学のデータということと，データ数が少ないことを考えると正規分布と仮定しても問題ないと思うよ．統計処理の結果は，前にも言ったと思うが，あくまでも科学的考察の補助として扱うべきなんだ．統計には，このように柔軟に考えてもいい部分と，厳密に考えないといけない部分があることを忘れないでほしい．

Aさん：統計って，思ったより柔軟なんですね．

X教授：正規分布を仮定しても問題ないといえる根拠として，中心極限定理というのがあるんだ．同じ母集団から無作為に抽出された標本の平均値は，標本数が大きくなると真の値に近づき，真の値との誤差は正規分布になることが保障されている．つまり，標本が正規分布から少し外れても，標本の平均値は正規分布としても問題ないと言える．

B君：わかったような，わからないような．標本の平均値は一つしかないのにその分布と言われても……

X教授：そこのところはまた後日詳しく説明するよ．と

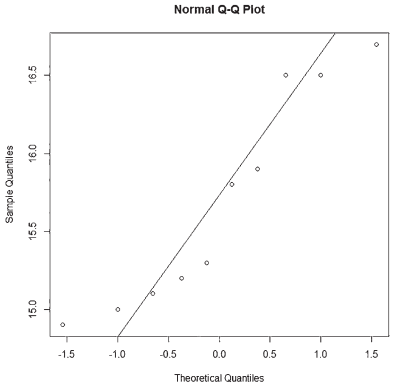

図2．AさんのQ-Qplot

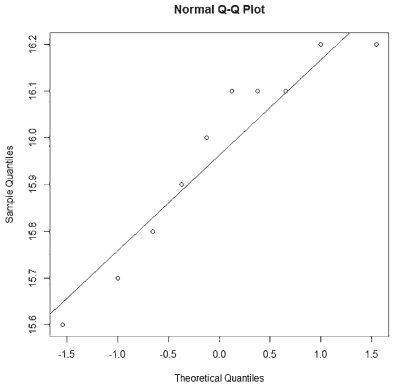

図3．B君のQ-Qplot

ころで, 標準偏差と標準誤差の違いは調べてみたかね.

B君：忘れてました.

Aさん：そうだと思ったので，調べてきました．標準偏差は，今回の実験データのようなものを1群の標本と言って，この1群の標本のばらつきを表す量です．標準誤差は，繰り返して実験を行った時の平均値の標準偏差で，平均値の精度を表す量です.

X教授：その通りと言いたいが，意味が分かっているかね．標準偏差がデータのばらつきを表すにはデータが正規分布に従うという条件が成り立つ必要がある．つまり，どんな場合でも単純に標準偏差を求めればいいわけではない．生物工学のデータなら，ほとんどの場合, 正規分布を仮定できるから, 問題はないと思うが. 標準偏差をSDとすると「平均 ± 1.96SD」の範囲に95%のデータが入ってくる．一方，標準誤差（SE）は母集団の平均が「平均 ± 1.96SE」の範囲に95%の確率で存在すると推定できるという意味になる．つまり，生物工学分野でデータのバラツキを表す場合は標準偏差を使うべきなんだ．分析方法の精度の良さを示したい場合なんかは標準誤差を使うべきだね．カッコいいから標準誤差を使うというのは, 前にも言ったが, 大間違いだな.

B君：反省します.

母平均の区間推定

母集団の平均「母平均」は直接測定することができないが，標本のデータからどの範囲の値かを推定することはできるというのが統計学の立場である．母平均の存在する範囲を推定することを「区間推定」と言う．区間推定を行う時にも，よく聞く言葉だと思うが「有意水準」を仮定することが必要になる.

X教授：まず，有意水準（α）を知っているかね.

B君：聞いたことはありますが，正確な意味はよく分かりません.

X教授：そうだと思うよ．有意水準とは，簡単に言えば「正しいことを間違っていると判定してしまう確率」のことなんだ. 昔は「危険率」とよばれたこともあった. もう少し，厳密な意味は，仮説検定を説明するときに話そう.

Aさん：今まで統計を勉強していたけど，機械的においていたので意味なんて考えたことはありませんでした. 少し，分かったような気がします.

表2．正規分布表の要

有意水準α	$z(\alpha/2)$
0.05	1.96
0.01	2.33

X教授：先程のデータを使って区間推定を行ってみよう．$\alpha = 0.05$ としたとき $100(1-\alpha)$ の確率で母平均が存在すると仮定できる区間を $100(1-\alpha)$% の信頼区間と言う．$\alpha = 0.05$ とすると，ちょうど95%信頼区間を出すことになるからね．さっき，母集団の平均が「平均 ± 1.96SE」の範囲に95%の確率で存在すると推定できると言ったが, これは, 厳密に言えば母集団の分散（母分散）が分かっている場合なんだ．母分散が標本の不偏分散と等しいことが分かっているときは，この式でいい.

まず，母分散が分かっている場合から始めよう．表2を見てほしい．ここに，標準正規分布表から区間推定によく使う数値が抜き出してあるんだ.

正規分布の中で平均が0，標準偏差が1の正規分布を標準正規分布と言って，いろいろな計算ではこの分布に合わせるようにするんだ．データが正規分布に従っているとすると，データを X，そのデータの平均を $\bar{X}$，不偏分散を σ^2 とすると $\dfrac{X-\bar{X}}{\sigma}$ は標準正規分布に従うことが知られている．これをデータの標準化と言うんだ.

母平均を μ，データ数を n とすると $\dfrac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sigma}$ が標準正規分布に従う．この値が標準正規分布の95%データの集まる区間にあればいいと考えるんだ．つまり，$-z\left(\dfrac{\alpha}{2}\right) < \dfrac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sigma} < z\left(\dfrac{\alpha}{2}\right)$ となる．$z\left(\dfrac{\alpha}{2}\right)$ を標準正規分布の切断点と言う.

図4を見てほしい．これは標準正規分布のグラフで平均（0）を中心に左右対称になっている．そして，信頼区間から外れる区間が左右にあることを矢印で示している．この矢印の区間に入る確率は同じになるから，左右で2.5%ずつ，合計5%が外れることになるんだ．こういう意味で $z(\alpha/2)$ と表されているんだ．上式を変形すると $\bar{X}-z\left(\dfrac{\alpha}{2}\right)\dfrac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z\left(\dfrac{\alpha}{2}\right)\dfrac{\sigma}{\sqrt{n}}$ となって，「平均 ± 1.96SE」になるわけだね.

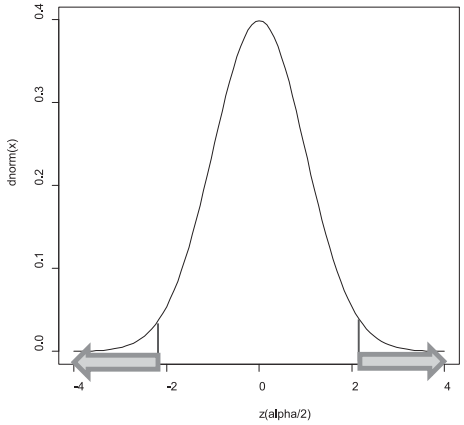

図4．標準正規分布（> curve(dnorm(x), -4, 4, xlab=”z(alpha/2)”）というRコマンドで作成）

Aさんのデータで95％信頼区間を出してみよう．

```
> mean(A)-1.96*sd(A)/sqrt(length(A))
[1] 15.26513
> mean(A)+1.96*sd(A)/sqrt(length(A))
[1] 16.11487
```

15.265から16.115の範囲になる．

では，母分散が分からない場合はどうなるかと言うと，標準正規分布の代わりにt-分布を使うんだ．t-分布については次回詳しく説明しようね．

母分散が分からない場合でも，不偏分散を使うことは変わりないんだが，$z\left(\frac{\alpha}{2}\right)$が使えないんで，これに代わる値が必要になるんだ．t-分布を使う場合は$z\left(\frac{\alpha}{2}\right)$の代わりに$t_{n-1}\left(\frac{\alpha}{2}\right)$を使う．$n-1$は自由度と言う値で情報量と関係がある．値としては（データ数–1）なんだが意味は違うんだ．詳しいことは次回に回すね．今回二人ともデータ数が10なので，自由度は9になる．表3のt-分布表を見ると自由度9で上側確立2.5％の値は2.262になる．

$$\bar{X}-t_{n-1}\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}<\mu<\bar{X}+t_{n-1}\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}$$

のそれぞれの値を入れればいい．

Aさんのデータでいうと

```
> mean(A)-2.262*sd(A)/sqrt(length(A))
[1] 15.19967
> mean(A)+2.262*sd(A)/sqrt(length(A))
[1] 16.18033
```

15.200から16.180の間になる．分布が変わったのと，データ数が少ないことから少し広めの区間になる．

B君：データが少ないと，さっきからおっしゃっていますが，僕らの感覚から言うと10個のデータは多いと思うんですが．

Aさん：10個のデータをとるのは大変ですよ．

X教授：実験を行う立場からするとそうだと思うよ．でもね，統計学的には少ないんだ．t-分布を使おうとすると最低でも6個のサンプルは必要と言われているし，正規分布だと3ケタくらいは必要になる．物理では，何百回も測定を繰り返すが，意味が分かるだろう．生物では繰り返すことが難しかったり，労力が大変なので3回としているようなんだが，統計的には少なすぎると言わざるを得ない．このことを頭において，統計処理結果を慎重に扱うという前提で3回分のデータの統計処理で議論していると理解したらいいんじゃないかな．

Aさん：健康診断の血液検査は1回の測定ですよ．

X教授：血液検査は，これまでに多くのデータの蓄積があるし，その年に多くのサンプルも集まる．これらのデータ集団を使うと異常値を検出できるんだ．次は，異常値の見つけ方と，実際に統計処理する場合，何回実験を行うかを考えてみよう．

B君：まだ，当分続きますね．

Aさん：楽しみです．

表3．t分布表の要約

自由度$n-1$	$t_{n-1}(\alpha/2)$，（$\alpha=0.05$のとき）
1	12.70
2	4.30
3	3.18
4	2.78
5	2.57
6	2.44
7	2.37
8	2.31
9	2.26
10	2.23

参考文献

第1回の参考文献1～5を参照のこと．

（【第3回】は94巻8号に掲載予定です．）