間違いから学ぶ

実践統計解析【第1回】

平均値にご注意を

川瀬　雅也[1]*・松田　史生[2]

この連載の目的は「生物工学分野の研究成果報告のデータ処理にありがちな誤りを他山の石として，統計処理法の理論的背景をおさらいする」ことである．統計処理は，実験結果の科学的評価に必須のツールとなっているが，独特の概念に基づくため，誤って使ってしまうことも多々ある．そして，統計ツールの正しい使い方を求めて，統計学の教科書をひもといても，母集団，信頼区間，確率分布，有意水準といった不可思議な概念の壁の前で呆然とするのみである．そこで，統計学を一通り学習した生物工学系研究者が統計ツールを活用する一助とすべく本連載が企画された．より実践的にするために，生物工学系の研究室で卒業研究を始めたばかりのAさんと研究室の先輩で院生のBさんと一緒にデータの処理法を，X教授から学ぶという形式をとる．

計算はPCに任せる

数理統計学の先生に見せれば，烈火のごとくお怒りになるタイトルかと思う．現実を見れば，データの統計処理を手計算でやっている方は皆無と思われる．この連載では“R”というフリーソフトを用いる[1]．Rについては多様な書籍[2]やwebサイト[3,4]が存在しており，適宜そちらを参照していただきたい．ここでは，すでに，皆さんのPCにRがインストールされているとして話を進める．詳しいインストール方法は上記の書籍やwebサイトを参照いただきたい．

データ処理事始め

卒業研究の最初の実験として，Aさんは大腸菌のコンピテントセルを作成することになった．指導教官は，その教育係として先輩のB君をあて，2人で並行してコンピテントセルを作成し，空ベクターでの形質転換効率を比較するように指示した．かっこいいところを見せたいB君は負けるわけにはいかない．Aさんも1日も早く一人前になろうと必死だ．2人は同じストックから大腸菌をそれぞれ培養し，同一のプロトコルでコンピテントセルを作成した．2人とも，3つに小分けしたコンピテントセルに同じ空ベクターを導入し，3枚の選抜培地プレートで一晩培養した．翌朝，見事に形質転換体のコロニーが観察された．その数は表1のようになった．

表1．コロニー数

Aさんのコンピテントセル	B君のコンピテントセル
15, 19, 22	15, 19, 28

B君：2人のデータの平均をとってみようか．
Aさん：私と先輩の作ったコンピテントセルのプレート当たりのコロニー数の平均値は，それぞれ，18.7個と20.7個と先輩の方が多いですね．
B君：（よっしゃー．先輩のメンツが立った！）
Aさん：けど先輩！3枚のうち2人の違いは3枚目のプレートだけなんですけど，これだけで差があると言っていいですか？統計学の講義では，平均がデータの代表値として使えない場合もあると習ったんですが，この場合はどうなんでしょうか？あと，ばらつきの代表値の標準偏差などは考えなくていいんですかね？
B君：え…あ，そうそう，うーんと…

向学心に燃えるAさんは，頼りにならない先輩にあきれるでもなく，叔父でデータ分析が専門のX教授（同じ大学に勤める）に相談しようと思い立ち，2人はX教授の研究室を訪ねた．

データを読み解く[5]

AさんはX教授への挨拶もそこそこに，B君を紹介しこれまでの事情を説明した．

X教授：なるほど．3反復の実験から得たデータの平均をとったわけだね．でもこの結果からB君作成コンピテントセルが優れていると評価できるかな？
Aさん：そうなんです．B先輩とまったく同じ操作をしたのに結果に差があるというのは納得いきません．
X教授：そうだねぇ．でも，そもそもどうして同じ実験操作をしたのに，3反復で結果が異なったのだろう？
B君：実験誤差というやつですか？
X教授：それそれ．同じ実験をしたつもりでも実験操作の微妙な差によって，結果にばらつきが生じてしまうんだね．そうだねぇ．いい機会だからもうちょっとデータを追加してよく考えてみようか．

*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）

X教授は，友人でもある2人の指導教官に了解を得た後，同じ実験を20反復で行うよう2人に勧めた．B君は「20反復の実験なんて普通しないよ…」とぶつぶつ言っていたが，がぜんやる気のAさんに押し切られる形で，実験室に戻って再実験を行い，翌朝表2の結果を得た．

表2．再実験のコロニー数

Aさんのコンピテントセル	B君のコンピテントセル
15, 13, 11, 13, 18, 19,	15, 17, 21, 23, 28, 19,
22, 21, 20, 16, 11, 16,	22, 24, 30, 26, 21, 19,
18, 10, 11, 12, 11, 14,	18, 20, 21, 22, 21, 24,
12, 11	22, 21

早速結果を持ってX教授のところに行くと，X教授はノートPCの‘R’を起動し，Rで平均をとる方法を説明しはじめた．Rのコンソールで，次のようにデータを入力する．

データ名を適当に付け（ここではAとB）“A <- c(　)”のカッコ内にデータを書き込めばいい．

```
> A <- c(15, 13, 11, 13, 18, 19, 22, 21, 20, 16, 11, 16, 18,
        10, 11, 12, 11, 14, 12, 11 )
> B <- c(15, 17, 21, 23, 28, 19, 22, 24, 30, 26, 21, 19, 18,
        20, 21, 22, 21, 24, 22, 21 )
```

平均はmean（データ名）で求めることができる．

```
> mean(A)
[1] 14.7
> mean(B)
[1] 21.7
```

Aさん：（ちょっとショック）そんな…どうしてなんでしょうか？

X教授：ややこしい話は後にして，まず，実験データの見方を紹介しよう．度数分布表とヒストグラムを知っているかね．

Aさん：習ったような気がしますが，覚えていません．

X教授：度数分布表とは，コロニーの数が17個だったプレートが何枚あったかという形でまとめた表のことだ．Rではこのようにすればいい．

```
> table(A)；度数分布表を書く関数
A；以下，上段が階級で下段が度数
 10  11  12  13  14  15  16  18  19  20  21  22
  1   5   2   2   1   1   2   2   1   1   1   1
> table(B)
B
 15  17  18  19  20  21  22  23  24  26  28  30
  1   1   1   2   1   5   3   1   2   1   1   1
```

この結果を見やすくグラフ化したものがヒストグラムになる．

Rではhist（データ名）でヒストグラムを書くことができる．

```
> hist(A)
> hist(B)
```

両方のヒストグラムを示しておく（図1，2）．Rで書かれたグラフを載せておくので，横軸は各々の度数分布表と対比させていただきたい．

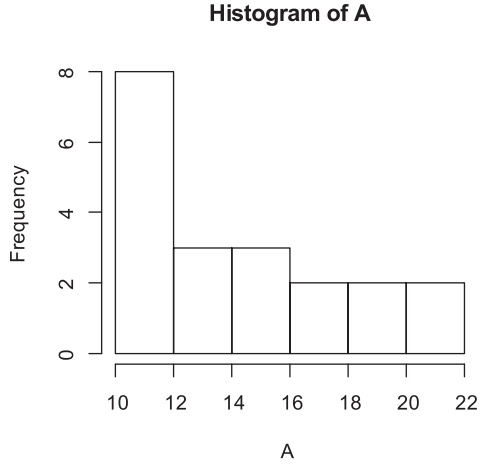

図1．Aさんのデータのヒストグラム

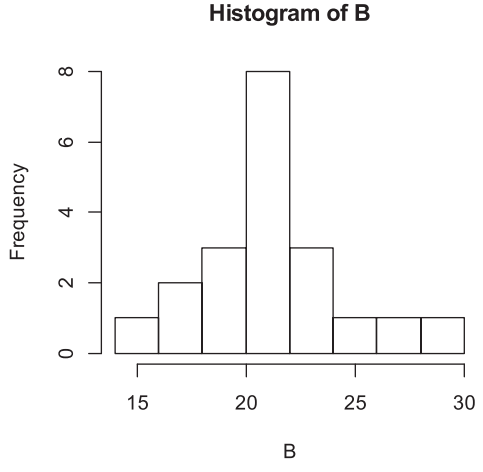

図2．Bさんのデータのヒストグラム

X教授：どうかね？
Aさん：先輩のデータは平均値にちかいプレートの数が多いけど，私のデータは右下がりで，全然違う形です．
X教授：このようにヒストグラムにするとデータの特徴がよくわかるんだ．統計的には2つのデータは分布が異なるみたいだね．次に四分位も見てみると平均の意味がよく分かると思うよ．

四分位とはデータを小さい値から順番に並べたとき，データの25%目の値を第1四分位，50%つまり真ん中の値を第2四分位（中央値やメジアンともいう），75%目の値を第3四分位というわけで，第1四分位と第3四分位の値の差を四分位範囲という．Rではsummaryコマンドを使う．

```
> summary(A)
   Min. 1st Qu.    Median    Mean 3rd Qu.    Max. 
   10.0    11.0    13.5    14.7    18.0    22.0 
> summary(B)
   Min. 1st Qu.    Median    Mean 3rd Qu.    Max. 
   15.00   19.75   21.00   21.70   23.25   30.00 
```

Quは四分位の意味であり，Medianは中央値である．
中央値はデータ数が偶数の場合2つの数がそれにあたるので，2つの数の平均を中央値とする．各四分位の値も同じように比例配分により小数になることもある．この結果を図にしたものが箱ひげ図（図3）である．

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

ヒストグラムと箱ひげ図は基本的に同じ情報を与えるもので，どちらか一方を使えばよい．箱の上端が第3四分位で下端が第1四分位である．箱から伸びたひげに当たる部分の下限が最小値，上限が最大値となるが，Rでは箱の上端もしくは下端から四分位範囲の1.5倍以上離れたデータを外れ値*として○で示すので注意されたい．また，箱の中にある太線は中央値を示している．

X教授：実験がうまくいっていることを確かめるには，データの分布が大事になってくるんだ．昨日，データ

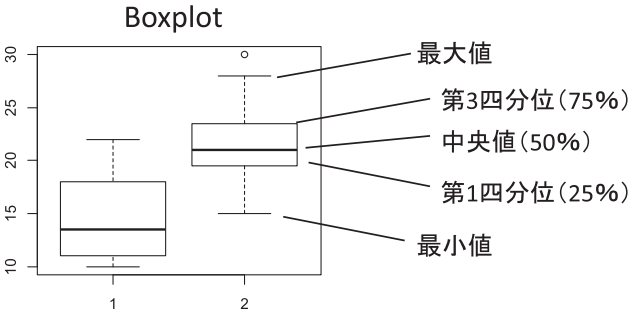

図3．Aさん（左）とBくんのデータ（右）の箱ひげ図

がばらつくのは実験誤差があるからだという話をしたけど，もし実験操作がうまくいっていれば，実験操作の各段階でランダムに生じたわずかな誤差が積算されて実験結果に反映されるはずだ．このランダムな実験誤差によっておきたばらつきは，中央値を中心にして均等に大小両方向に広がった，左右対称の釣鐘型に近くなるんだ．平均値は中央の度数最大のグループの階層にあると見てよい．このような場合は平均値を見ると「この部分にデータが集まっているのか」「データの分布の中心は」という感じで，データの特徴を平均で記述できるわけなんだ．
Bくんのデータがこれに当たるね．生物工学分野で行われる実験はきちんと行えば，こういう正規分布に近いデータが得られると考えていいんじゃないかな．
Aさん：私のデータは何か問題があるってことですか．
X教授：その通り．こういう場合の平均はデータ集団の特徴を表しているとは，とても言えない．そして，実験操作にランダムじゃない，系統的な誤差があったことを示している．実験操作がサンプル間で不均一になっていたと思うんだけど，なにか心当たりはあるかな．
Bくん：そうそう，混ぜるときにそっとするとか，温度とかちょっと気になってたところがあったんだよね．
Aさん:なるほど．帰ったらまた教えてください！先輩！
X教授：平均値の上手な使い方は，生物工学会誌でも以前紹介されているので[6]，読むと勉強になるよ．それから，データの分布が正規分布じゃないとき，中央値やもっとも出現頻度が高い値（最頻値；モード）が平均に代わって用いられることが時々あるから覚えておこう．

平均の話

Bくん：でも先生，JBBの論文で20反復の実験とか，中央値が載っている論文って僕は見たことありません．

*ここでは外れ値としているのはR上での外れ値の意味である．実際の外れ値の判定は難しく，回を改めて解説する．

3から5回反復した実験の平均値に，標準偏差のエラーバーが表記されているのが普通ですが，これって全部ダメ，ってことなんですか？

X教授：生物工学分野ではコストと労力の都合で，3回反復くらいしか実験ができないことが多いからねぇ．でも，さっきも言ったように，きちんと実施した実験から得られたデータは正規分布に近くなると仮定できる．この場合，実験データを，2種類の代表値（平均値と標準偏差）に実験の反復数をつけて，記載することは間違いどころかむしろ正しい．

Aさん：じゃあ私たちの実験も20反復もしなくてもよかったってことでしょうか？

X教授：いやいや，そうじゃないんだ．平均値と標準偏差と反復数だけでいいのは，データがきちんとした実験から得られて正規分布になると仮定できるとき「だけ」なので注意しよう．論文に出てくるようなデータは，そう仮定できるという暗黙の前提があるんだな．けれど，その仮定が怪しいんじゃないか，という日々の研究で出てくる今回のようなケースでは，データが本当に正規分布になっていることを確かめる必要があるけど，それには，3反復の実験データでは足りないんだ．ヒストグラムで分布を調べるためには，最低でも20反復くらいないとわからないだろ．

Aさん：けど，毎回20反復の実験は大変ですよね．

X教授：毎回確認する必要はないよね．新しく実験系を立ち上げたとき，初めてやる実験のときに多数反復して，正規分布に近くなることを1回確認しておけば，以降は仮定でいいんじゃないかな．

B君：なるほど，実験データがおかしいときなんかにも多数反復して正規分布になるか確かめれば，問題点を突き止めるのにいいかもしれないですね．

平均値の比較？

X教授：何となく，感じはつかめたかね．

Aさん：はい．やはり，平均と標準偏差が大事なんですね．

X教授：平均と標準偏差は，正規分布の形を決める重要な量になる．生物工学分野で得られるデータは，正規分布に近い場合が多いので，これをまず計算するんだ．

Bさん：Rで平均の計算法は教えてもらったんですが，標準偏差はどうすればいいんですか．

X教授：標準偏差は不偏分散から計算できる．

$$\text{不偏分散：}\ \sigma^2 = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2$$

この量は不偏分散とよばれる量で，$\bar{x}$は平均，x_iは個々のデータ，nはデータ数を表している．また，標準偏差は分散の平方根をとったもの（正の値）となる．

Rは関数電卓のように数式を入れても計算ができるが，ちゃんと分散や標準偏差を計算する関数が用意されている．先ほどのB君のデータを使って計算してみよう．

```
> var(B)      ；不偏分散を求める関数
[1] 12.64211
> sd(B)       ；不偏分散から標準偏差を求める関数
[1] 3.555574
```

また，繰り返して実験を行った場合，各回の平均は当然ながら同じではない．この平均の変動の様子を表すために標準誤差（平均の標準偏差）を求めることもある．

```
> sqrt(var(B)/length(B))      ；√(σ²/n)
[1] 0.7950505
> length(B)       ；Bのデータ数を求める関数
[1] 20
```

B君：標準誤差って聞いたことあります．標準偏差よりもエラーバーが短くなるから，データに有意差があるっぽく見えるって誰かが言ってました．

X教授：むむ．それは聞き捨てならないな．標準偏差と標準誤差では，意味がまったく違うんだ．それから，最初相談に来たとき，平均値が大きいから，いいコンピテントセルだと言ってたけどこれも正しい統計の使い方じゃない．次回は，その理由を正規分布と母集団から説明するから，覚悟しておいで．

Aさん（うれしそうに），B君（ちょっとびびり気味で）：よろしくお願いしまーす．

文　献

1) R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/
2) 船尾暢男：R-Tips―データ解析環境Rの基本技・グラフィック活用集，オーム社 (2009).
3) http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html
4) http://www.okada.jp.org/RWiki/
5) 統計検定のサイト（http://www.toukei-kentei.jp/）からリンクの情報や，標準教科書を参照いただきたい．
6) 川瀬雅也：生物工学，**91**, 4, 205 (2013).

（【第2回】は94巻6号に掲載予定です．）