間違いから学ぶ

実践統計解析【第3回】

データ数はいくつ必要

川瀬　雅也[1]*・松田　史生[2]

実験を行ってデータを取得し，統計処理する時，皆さんは何反復の実験，すなわちデータ数いくつ（n数とよく言う）で行うだろうか．ほとんどの方が何の疑いもなく“3”と答えるのではないだろうか．では，なぜ，“3”なのかという問いに答えることができるだろうか．今回は，この問いを考えてみたい．

データ数の疑問

Aさん：B先輩．なぜ，実験は3回繰り返さないといけないんですか？
B君：データを統計的に処理するためだよ．
Aさん：3回だけでいいんですね．
B君：そうだけど．
Aさん：なぜ，3回だけでいいんですか？
B君：……．先生が何があっても実験は3回反復って口を酸っぱくして言っていたから．
Aさん：？？？．先輩，理由を知らないんですか？

皆さんの研究室でも，もしかするとAさんとB君のようなやり取りがあるのではないだろうか．2人は，例によって例のごとく，再びX教授のもとを訪ねてきた．

X教授：いらっしゃい，待っていたよ．
B君：お手柔らかにお願いします．
Aさん：実は，……，と言う訳なんです．
X教授:なるほど．だがな，B君のような学生は，きっと，どの研究室にも多いと思うな．教えている教員も怪しいかもしれないな．
B君：そうですよね．みんな，知りませんよね．
X教授：調子にのるな．データ数がいかに重要かは，これまで言ってきただろう．統計処理を行ううえで，データ数がいかに重要かを，じっくり説明しよう．
Aさん・B君：よろしくお願いします．
X教授：まず，先生は，どう言っていたのかな？
B君：データの検定には，少なくとも3個のデータが必要なので，3個のデータをとるために“実験は3回行うこと”です．実際には，時間的なことを考えて3回の繰り返しでいつも終わっています．
X教授：なるほど．検定については，こちらも時間の関係で次回に説明するとして，なぜ，3回でいいかということから考えてみようか．

この図を見てもらおうか（図1）．これは，乱数を使って人工的にデータを作り，その分布を見たものだが，どうかな？

Aさん：データ数が3のときはどんな分布かさっぱりわ

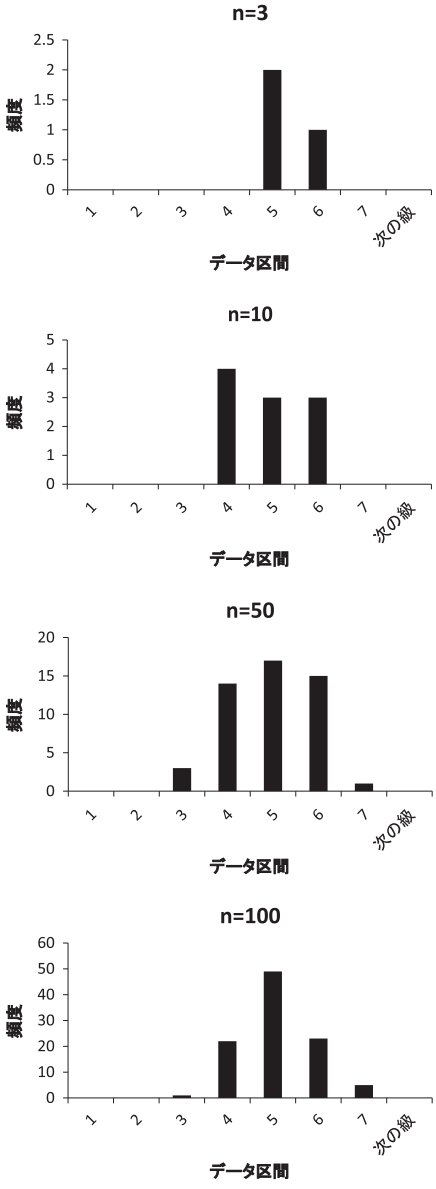

図1．乱数で生成したデータ数と値の分布

*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）

かりませんが，50になると何となく分かってきますし，100になると正規分布に見えてきます．

B君：僕もそう思う．

X教授：そうだろう．まったくどんな分布をしているのか分からない場合，100くらいのデータがないと分布を知ることはできそうにないことが分かるだろう．しかし，100回同じ実験を繰り返すなんて言うことは無理だと思うな．研究費も限りがあるし，時間も掛かるので，卒業などを控えた学生には絶対に無理だ．だから，最少の繰り返し回数がどのくらいかが重要になるわけだ．

Aさん：この図を見ると3回でも少なそうに思います．

B君：でも，うちの研究室だけが3回と言う訳ではなく，学会なんかに行くと，どの研究室も3回だよ．「赤信号，みんなで渡れば怖くない」ですか．

Aさん：何ですか，それ？？？？

X教授：えらい古いギャグを知っているな．話をもとに戻すが，B君の言う通りだと思う．みんなが3回だから，自分たちも3回でいい，まさに「赤信号，みんなで渡れば怖くない」だな．

B君：ほら！正解だろう．

Aさん：今日は嵐が来そうですね．

X教授：面白いから，もっと聞いていたい気もするが，先に進むとしよう．まず，データが一つではダメなことはわかるね．とんでもない失敗データでも一つしかないと分からないからね．

　では，二つだとなぜダメなのか，自由度を使った説明が多い．二つのデータから平均値を計算すると，残りの自由度は1となる．データにバラツキがないから統計的に意味のある分散は計算できない．というのが直感的な説明かな．

Aさん：最低三つのデータでいいなら，実験は3回でいいんじゃないですか．どうして3回だけだと不十分なんですか？

X教授：もう一つの説明は，平均値の95%信頼区間を計算してみるというものだ．前回，B君が酵素活性を10回測定したデータを使って計算してみよう．

```
>B <- c(15.7,16.1,15.6,16.0,16.1,15.9,16.2,16.1,15.8,16.2)
> t.test(B)
```

を実行した出力の

```
95 percent confidence interval:
 15.819 16.121
```

の部分が95%信頼区間を示している．この結果から活性値の真の値（母平均）は15.8から16.1の間にある確率は95%といえる．

Aさん：かなり狭い範囲ですね．これはB先輩の実験のばらつきが小さかったからですか？

X教授：では，実験の反復数が5回，3回，2回として計算してみよう．

```
> B <- c(15.7,16.1,15.6,16.0,16.1)
> t.test(B)
95 percent confidence interval:
 15.6088 16.1912 #5反復のとき15.6から16.2
> B <- c(15.7,16.1,15.6)
> t.test(B)
95 percent confidence interval:
 15.14276 16.45724 #3反復のとき15.1 から16.5
> B <- c(15.7,16.1)
> t.test(B)
95 percent confidence interval:
 13.35876 18.44124 #2反復のとき13.4 から18.4
```

B君：反復数が減ると95%信頼区間がどんどん広がってますね．2反復では統計的な意味もなく，仮に計算した信頼区間も広すぎだ．

Aさん：やはり最低3反復，できれば5反復くらいデータが必要ですね．じゃあ何点のデータがあれば十分なんでしょう？

X教授：仮説検定を例にするのが分かりやすいので，簡単な例で大まかなところを説明する．詳しい説明は次回にまわすので，楽しみにしておいてくれるかね．

B君：とうとう本番ですね．僕も，検定には悩まされているんです．

X教授：君たちがよく使うのは「平均の差の検定」という種類の検定だと思う．たとえば，2種類の微生物AとBのどちらが高い抗生物質の生産能力を持っているのかを調べようとしているとしよう．

A：12.5，13.2，13.3

B：11.1，10.8，11.4

例なので単位は気にしないことにするが，上記のようなデータが得られたとする．この時使われるのが「平均の差の検定で，多分，その中でもスチューデントのt-検定を使うと思う．

B君：僕も，いつも使っています．

X教授：いつもと言うのは感心しないが，理由は次回話すことにして，今はスチューデントのt-検定を使ってみると，

```
> A <- c(12.5,13.2,13.3)
> B <- c(11.1,10.8,11.4)
> t.test(A,B,var.equal=T)
   t = 6.2192, df = 4, p-value = 0.003403
```

となる．

前回，有意水準（α）という言葉が出てきたのを覚えているね．通常の検定では，αを0.05に設定することが多い．この場合，p-value ＜ 0.05なので「違いがあると見てもよい」という結果になる（次回に正しい意味を説明する）．一方，有意水準は簡単に言うと「正しいことを間違っていると判定してしまう確率」（ここでは「本当は違いがないのに，間違って違いがあると判定してしまう確率」）と説明していたと思う．

Aさん・B君：その通りです．

X教授：統計学的には「正しいことを間違っていると判定する」間違いを“第1種の過誤”と言うんだ．この第1種の過誤の確率が有意水準（α）ということになるわけだ．

Aさん：そうしたら逆の「間違っているのに正しい」とする間違いもあるんですよね？

X教授:その通り．なかなか鋭い．Aさんの言う「間違っているのに正しい」とする（ここでは，本当は違いがあるのに間違って違いがないと判定してしまう）間違いを第2種の過誤と言うんだ．第2種の過誤の確率をβとして（1 − β）を検出力と言うんだ．違いを見つける能力の大きさとでも言えばいいのだろうね．

統計的にデータを考えるという場合，多くの研究者は有意水準にしか注目しない．だから，データ数の大切さに気が付かないと言っていいと思う．これは，今の統計教育の落とし穴だと思うな．統計教育では検定の方法については訓練するが，データ数については何も教えていないからな．

B君：検定法の演習のときは，問題のデータで計算するだけで，データ数が適切かどうかなんてまったく気にしなかったですし，説明もなかったです．

X教授：検出力まで勉強する講義はないと思うので，この際，勉強しようか．

データAとBを使って検出力分析を行ってみると，

```
> mean(A)-mean(B)
[1] 1.9
> sqrt((2*var(A)+2*var(B))/4)
[1] 0.3741657
> power.t.test(n=3,d=1.9,sd=0.374)
     Two-sample t test power calculation
              n = 3
          delta = 1.9
             sd = 0.374
      sig.level = 0.05
          power = 0.993979
    alternative = two.sided
```

となる．ここで，powerの値，検出力の値に注目してほしいんだ．この例では約0.994となっているね．検出力は0.8を超えることが望ましいとされている．今回は0.8を超えているから，データ数が3個でも議論に使ってもいいということになるわけだ．つまり，この例で使った実験系だと，君たちが何時もやっているように，3回の実験で十分となる．

別の例を見てみよう．同じ実験をCとDという微生物で行ったとしよう．結果として

C：125，132，133

D：121，118，133

というデータが得られたとする．同じように検定を行ってみると，

```
> C <- c(125,132,133)
> D <- c(121,118,124)
> t.test(C,D,var.equal=T)
   t = 2.9459, df = 4, p-value = 0.04214
```

となって，やはり，違いがあるとみていいという結果になる．検出力分析を行うと，

```
> mean(C)-mean(D)
[1] 9
> sqrt((2*var(C)+2*var(D))/4)
[1] 3.741657
> power.t.test(n=3,d=9,sd=3.74)
     Two-sample t test power calculation
              n = 3
          delta = 9
             sd = 3.74
      sig.level = 0.05
          power = 0.6047886
    alternative = two.sided
```

で，検出力不足になっている．つまりこの実験系では，データは3個では不足ということになるんだ．いくつのデータが必要かと言うと，

```
> power.t.test(d=9,sd=3.74,power=0.8)
     Two-sample t test power calculation
```

```
              n = 3.945541
          delta = 9
              sd = 3.74
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

4個ずつデータが必要という結果になる．もし，データのバラツキが大きく標準偏差（sd）が大きな場合は，

（sd=9の場合）

```
> power.t.test(d=9,sd=9,power=0.8)
     Two-sample t test power calculation
              n = 16.71477
          delta = 9
              sd = 9
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

データは17個ずつ必要になるわけだ．

検定の話をしていないので，分からないところもあると思うが，「必要なデータ数は得られたデータによって違ってくる」ということを，今は分かってくれればいい．どうかな．

B君：何となく，分かりました．

Aさん：いつも，検出力分析をして確認しないといけなんですか．

X教授：そこが，実は大事なところなんだ．統計学的には，いつも，検出力分析を行うということになるのだが，君たちは科学の世界にいるわけだね．

Aさん・B君：そうです．

X教授：今見せたように，統計的にデータを解釈するため検定などの方法を使って計算することを統計処理と言うことがある．科学の世界の人は，統計処理の結果が絶対だと間違った認識でいることがよくある．あくまでも統計処理の結果は「この統計処理法で計算したときの計算結果にすぎない」ことを忘れないでほしい．統計処理の結果は，その場のデータだけで判断をしているだけで，その背景にある，これまで積み上げられてきた事実などはまったく無視しているわけだ．先程の例でも，両方とも違いがあると見ていいが，一方は検出力不足であるとの結果が出た．もし，これまで，多くの実験がCとDについて行われ蓄積がある場合，つまり，CとDの性質がある程度分かっている場合，その蓄積から考えて，3個のデータの比較の結果が妥当であると科学的に言えるなら3個のデータの比較でもいいと言えると思うがね．

B君：なるほど．

Aさん：過去に蓄積がなく，初めての場合はどうですか．

X教授：その時は，先の例で見せたように検出力を参考にして，必要なデータ数を出す必要があるのではないかな．最初は，面倒に思わず実験を繰り返すことだ．

蛇足かもしれないが，統計処理結果はあくまでも計算結果だ．仮に，統計的に「差があると言えない」と出た場合でも，完全に差がないと言っているわけではない．科学的に差があると考えるべき場合は，データをもっととるとか，検討の方向を変えるなどの工夫が必要になる．

B君：今までは，差がありそうに思えても，検定で差がなさそうなら見込みがないということになっていたんですが，検定結果だけで結論を出すのは早いわけですね．

X教授：そういうことだね．統計処理の結果は，少し古いかもしれないが「水戸黄門の印籠」ではない．

統計処理は補助手段であり，統計で絶対的な結論を出すことはできないんだ．逆に，差があると出ても注意が必要になる．よく統計的に差があると言えるから大丈夫だと思う人がいるが決してそうではない．その気になれば，ある程度違いのありそうなデータなら検定で差があるという結果を出すことができる．こうなると，もはや科学ではなく数字遊びになってしまう．また，こんなことをやって，論文を書いたら「不正な論文」となってしまう．上で話したように，統計は計算に掛けているデータしか見ていないから，どんなデータを計算しているかも大事になってくる．

ところで，「外れ値」という言葉を知っているかな．

Aさん・B君：知っています．

X教授：「外れ値」が曲者なんだ．検定の話の後に，「外れ値」も考えてみよう．

参考文献

1) 川瀬雅也，松田史生：生物工学，**94**, 208 (2016).

（【第4回】は94巻10号に掲載予定です）