間違いから学ぶ

実践統計解析【第9回】

相関と相関係数

川瀬　雅也[1]*・松田　史生[2]

データに相関がある，あるいはないという議論をよく耳にするが，よく理解して使わないととんでもない間違いを起こすことがある．例の二人は大丈夫だろうか？

相関がある？ない？

Aさん：先輩，今日の研究室セミナーは准教授の先生にこてんぱてんにやっつけられてましたね．

B君：とほほほ（涙）．大腸菌中の遊離アミノ酸組成の制御機構を調べようということで，10変異株を培養してアミノ酸含量を頑張って測定したんだ．そうしたらロイシンとイソロイシン含量の間に相関が見られたんだよね（図1）．それで，ロイシン含量がイソロイシン生合成に何らかの影響を与えている可能性がある．と考察したら……あああ，その後のことは思い出したくないよ．でも，どこがそんなにまずかったんだろ．

Aさん：先輩，アイスをおごってあげますから，気を落とさないでくださいね．まずはX教授に相談しましょ．

X教授のもとを訪ねた二人はこれまでの経緯を説明した．

X教授：まずはデータを見せてみなよ．（図1を見て）ほほぉ，それでB君はどうしたのかな？

B君：Rのcor.test関数を使って相関係数を計算したら$r = 0.76$になりました（下記結果の一番最後）．

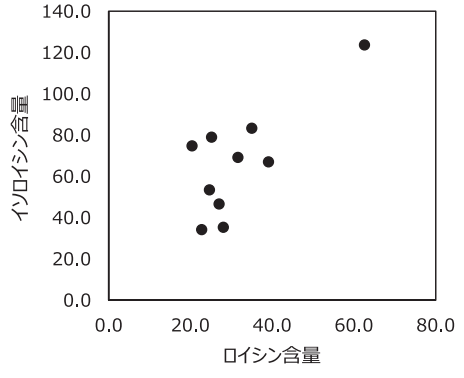

図1．B君がセミナーで発表したデータ（単位は略）

```
> Leu <-c(28.1, 35.0, 24.7, 31.6, 39.1, 22.7, 62.6, 25.2, 27.0, 20.4)
> Ile <-c(35.3, 83.3, 53.4, 69.1, 67.0, 34.2, 123.6, 79.0, 46.6, 74.8)
> cor.test(Leu, Ile)
        Pearson's product-moment correlation
data:  Leu and Ile
t = 3.3263, df = 8, p-value = 0.01044
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2540396 0.9403795
sample estimates:
      cor
0.7618138
```

B君：この結果から有意水準αを0.05としたときに相関係数がゼロという帰無仮説を棄却できるので（p-value = 0.010），有意な相関であると考えました．また，統計の講義でならった表1を見て「強い相関がある」と結論づけました．それで，ロイシン含量がイソロイシン生合成に影響しているんじゃないかと，議論しました．

表1．相関係数と相関の強さ（複合同順）[1)]

相関係数	相関の強さ
0.0～± 0.2	ほとんど相関がない
± 0.2～± 0.4	やや相関がある
± 0.4～± 0.7	かなり相関がある
± 0.7～± 1.0	強い相関がある
特に，± 0.9～± 1.0	非常に強い相関があるという場合もある

X教授：うーん．惜しいだけに，准教授の先生が怒り出しても仕方ないかもね．相関係数や相関については，統計学の授業でも回帰分析の前座的な扱いしか教えないから，詳しいことを知らない学生も多いと思う．今回は詳しく話をしていこうか．まず，Rのcor, cor.test関数，Excelのcorrel関数はデフォルト設定では

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）

Peasonの積率相関係数というものを計算している．データX，YのXのi番目の値をx_i，Yのi番目の値をy_i，Xの平均を$\bar{x}$，Yの平均を$\bar{y}$とすると，Peasonの積率相関係数rは

$$r=\frac{\frac{1}{n}\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_i(x_i-\bar{x})^2\cdot\frac{1}{n}\sum_i(y_i-\bar{y})^2}}$$

となるんだ．分母は標本分散の積の平方根，分子は共分散に当たる．これが，統計学の教科書に載っている相関係数だ．nはデータ数だね．

Aさん：共分散というのは初めて聞きました．

X教授：共分散はXとYがどの程度似ているかを表している量と考えていい．それからPeasonの積率相関係数はデータの母集団が正規分布の時に，相関の程度を表すことができるんだ．

Aさん：相関係数を求めるときもデータが正規分布になっていないといけないんですか？

B君：なので，今回のデータで正規性の検定を行っているのですが（Kolmogorov-Smirnov検定，連載第4回[2]参照），どちらも正規分布にしたがってないとはいえない，という結果になっているんです．

Aさん：でも准教授の先生は，相関係数を大きく見積もりすぎじゃないのか？ってコメントしてましたよね．どういうことなんでしょう？

X教授：データを見てみると1点だけ他からかなり外れたデータがあるよね．そのデータを除いた9点で散布図を書いてみたらどうなる？

Aさん：なんだか，相関があるとは言えなくなっちゃいましたね（図2）．

B君：この9点のデータで計算をやり直すと，相関係数はr = 0.31，p-value = 0.4176で検定もパスしないぞ，あれれ．それでも，表1によれば「やや相関がある」と言えるのじゃないでしょうか？

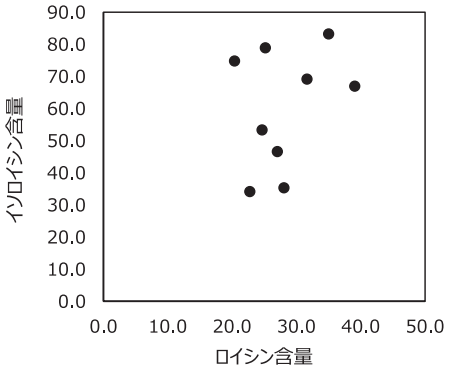

図2．図1からデータを一つ除いたグラフ（単位は略）

X教授：心理学とか医療分野では50とか，数百個というデータ点から相関係数を計算する．そんなときは表1を用いた解釈が妥当かなと思うよ．けど，今回の場合はデータの点数が少なすぎる．たかだか10点のデータでは，データの母集団が正規分布に従うかもはっきりしない，それに，外れ値が一つあるとそれに引っ張られて，相関係数が大きく計算されてしまう場合がある．データ点数が少ないときは，かなり慎重になる必要があるんだよ．さっき説明してくれたB君の解析法に間違いはないんだ．けど，実験結果から，本当に相関があると自信を持っていえる？

B君：……確かに，1点データを除くと結果がひっくり返るようではあやしいですね．准教授の先生はそこのところを指摘したかったんですね．

Spearmanの順位相関係数

Aさん：じゃあどうするのがいいのでしょうか？

X教授：データ点数が少ない場合や，データの母集団が正規分布に従うことが確実ではない場合は，Spearmanの順位相関係数を使うんだ．Spearmanの順位相関係数では，まず各データの値をデータ内の順位に変換し，順位のデータでPeasonの積率相関係数を計算する．Rだと簡単に計算できるよ．さっきのcor.testにmethod="s"を指定するといいんだ．

```
> cor.test(Leu, Ile,method="s")
        Spearman's rank correlation rho
data:  Leu and Ile
S = 92, p-value = 0.2042
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4424242
```

Aさん：同じデータなのに，相関係数r = 0.44になりましたね．この値から相関はある．としていいのでしょうか？

X教授：まずバイオ分野でも相関係数を計算するなら最低10点か15点くらいはデータ点数が必要だと言われている．今回はぎりぎりセーフってとこだね．そのうえで相関係数は，$|r| > 0.6$か$|r| > 0.7$のときに相関ありとする場合が多いかなぁ（実はあんまり根拠はないんだけどね）．Spearmanの順位相関係数は外れ値に強く，より保守的に相関を評価しているといえる．だから，Spearmanの順位相関係数を使うと論文の査読の

時にもケチをつけられにくくなるかもしれないね.

Aさん：でも，B先輩のデータからロイシンとイソロイシン含量の間に相関があるとは言いにくくなっちゃいますね.

X教授：ただ，相関がないとも言いきれないよね．弱い相関を議論したい場合は，やはりデータの点数を増やすのがいいと思うよ.

相関関係と因果関係

X教授：それからもう一つ，准教授の先生の指導が厳しくなったのは，「ロイシン含量がイソロイシン生合成に影響している」という考察のほうが原因じゃないかな．B君は偽相関って聞いたことあるかい？

B君：統計の講義でちらっと聞きました．えーっと．たしか，朝食をきちんと食べる習慣のある学生ほど，テストの正答率が高いという相関についてでした（http://www.maff.go.jp/j/seisan/kakou/mezamasi/about/databox.html）.

Aさん：だから朝食を食べると成績が良くなるというわけですね．つまり朝食を食べると，朝からブドウ糖がエネルギー源になって脳が働くんでしょうか？

X教授：もし，その説が本当なら，朝食を食べる習慣とテストの正答率の間の因果関係が，相関として観察されたことになるね．でも，こうとも考えられないかな．子供に毎日勉強する，朝ご飯を食べる生活習慣をきちんと付けさせている親の子供は，朝ご飯を食べる習慣があるし，テストの正答率も上がる．もし，この説が本当なら，朝食を食べる習慣とテストの正答率に観察された相関は，とくに両者の因果関係を反映したものではないので，偽相関という.

B君：なるほど，二つのデータに相関があっても，それだけで因果関係がある根拠にはなりませんよね.

X教授：それから，仮に因果関係があったとしてもどっちが原因で，結果なのかという，因果関係の向きまではわからないよね．B君はこのデータから「ロイシン含量がイソロイシン生合成に影響している」という仮説を立てたんだけど，ひょっとしてたまたまロイシン含量をX軸に，イソロイシン含量をY軸にプロットしたからじゃないのかね.

B君：（小声で）すみません．そのとおりです.

X教授：准教授の先生は相関があったという結果から，オーバーディスカッションをしてしまったところを戒めたかったんだろうね.

Aさん：でも教授，調べたんですが，大腸菌のロイシンとイソロイシン生合成経路の最終ステップの反応は同じ酵素が触媒します．この知見から，ロイシン含量とイソロイシン含量が相関するはず，という仮説を立てて，今回の見られた弱い相関は，この仮説を支持するものである．というような議論をすることは，ムリなんでしょうか？

X教授：そんなことはないよ，ただ観察された相関が弱いから，他の証拠もないと説得力には欠けるね.

B君：相関係数も奥が深いですね．これまで，相関係数の値の大きさだけ気にして，相関係数の計算法や解釈には無頓着だったので，これからほんと気をつけます.

Aさん：ところで，Excelでグラフに近似線を入れたときR^2という値が出るのですが，これも相関係数ですか.

X教授：これは“決定係数”とよばれる量だ．よく，相関係数の2乗であり正の値だけしか出ないと書かれているけど，実際には定義が複数あって定まっていないし，単に相関係数を2乗したものではないんだ．定義によっては負の値をとるときもある．Aさんの言う，Excelの近似線は回帰分析の結果なので，決定係数の話も含めて次回は，回帰分析の話をしようか.

統計的感覚

前回出したクイズの問題は，次のようなものであった.

Zさんが，細菌Jの画期的な簡易検出法を開発した．細菌Jは，毒性が強いため注意が必要な細菌であるが，これまで感度の良い検出法がなかったので，検査場所でサンプリングしたDNAをPCRで検査して，細菌Jの有無の判定を行ってきた.

Zさんの方法では，細菌Jが一つでも存在する場合，98％の確率で検出が可能である．細菌Jが全く存在しない場合，間違って存在するとしてしまう確率は5％である．実際に，細菌Jが検査する場所に存在する確率は経験的に4％であるとされている.

ある場所を，Zさんの方法で検査したところ，細菌Jを検出した．どう考えればいいか.

さて，例の二人はどう答えるだろうか.

X教授：宿題の答えは出たかい.

B君：はい．細菌Jが一つでも存在すると98％検出可能ですから，検出されたからには90％くらいの確率で考えるべきだと思います.

Aさん：細菌Jが検査する場所に存在して正しく検出する確率は3.92％（= 0.04*0.98），存在しないのに，誤って検出する確率は4.8％（= 0.96*0.05）です．ということは検出した事例のうち，0.0392/(0.0392 + 0.048) = 0.4495で45％の確率で実際に菌が存在することになります（図3）.

B君：そんな低いの！

X教授：さすがAさんは惑わされないね．正解だよ．どうしても検出確率が高い方法となると，そちらに注目してしまって，実際はいないのに誤って検出してしまう確率を忘れてしまう．そうすると，感覚として「検出された」，すなわち「高い確率で存在する」と思ってしまう．全体をよく見て，ある特定の数字に惑わされないということが重要で，これが統計的な感覚というべきものかな．

　じゃあ，おまけで，この問題をベイズの定理に基づいて説明してみよう．

Aさん，B君：ベイズの定理？初めて聞きます．

X教授：これまで勉強してきた統計手法は，すべて母集団の分布（多くの場合，正規分布）を想定してきたね．言い方を変えると，実験中に起きた経験を一切考慮に入れていないと言ってもいいんだ．

Aさん：母集団からサンプリングを行う，という考え方ですよね．統計の授業では先生が口を酸っぱくしてその重要性を説明してくれたのですが．

X教授：一方，ベイズの定理に基づくベイズ統計学では事前に母集団があるという考え方をしない．その代わりに経験を取り入れて考えるんだ．まず，検査をする以前に「細菌Jが検査場所に存在するな」と考えている確率を事前確率$\Pr(B)$と呼ぶ．この場合問題文にあるように4%になる．$\Pr(A)$は細菌Jが検査で検出される確率としよう．それから，検査を実施し，Jが検出されたときに，細菌Jが実際に検査場所に存在する条件付確率をベイズ統計学では事後確率$\Pr(B|A)$と呼ぶ．$\Pr(A|B)$は細菌Jが検査場所に存在するときに検査で検出する条件付き確率（問題文より0.98）になる．これらの関係を示した式 $\Pr(B\,|\,A)=\dfrac{\Pr(A\,|\,B)\Pr(B)}{\Pr(A)}$ がベイズの定理だ．

B君：これだけですか．

X教授：ここからがすごく面白いから．事前確率$\Pr(B)$が4%のとき，Aさんの書いた図3のように，細菌Jが検査で検出される確率$\Pr(A)$は0.98*0.04 + 0.05*0.96 = 0.0872になる．そこで，検査を実施しJが検出されたとしよう．検査後に細菌Jが検査場所に存在する事後確率はベイズの定理から

$$\Pr(B\,|\,A)=\frac{0.98\times 0.04}{0.0872}=0.4495$$

と計算できるね．

B君：Aさんの議論と同じ結果ですね．

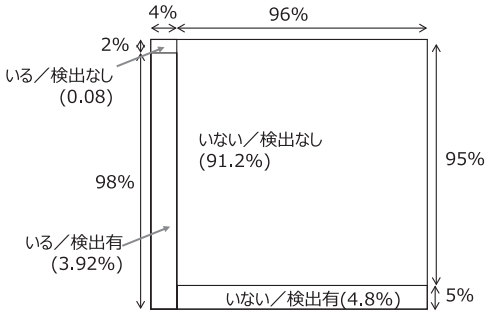

図3．Aさんの回答

X教授：じゃあ，もう一回検査をするとしようか．これまで勉強してきた統計手法だと，「細菌Jが検査場所に存在するな」と考える確率は4%で変化しない．同じ母集団からサンプリングするからだね．

　一方，ベイズ統計学では「検査の結果，細菌Jが検査場所に存在する確率が4%から45%に上がった．」というふうに考え，事前確率$\Pr(B)$を0.45に更新する．

Aさん，B君：は？？

X教授：つまり，問題文の検査する場所は一般的な場所のことなんだが，今回，検査した場所に限っては，そこの細菌Jが存在する事前確率$\Pr(B)$は0.45であると考えるんだ．経験値を生かすということだね．事後確率を計算すると$\Pr(B|A)$ = 0.94になる．つまり，細菌Jが今回の検査場所に存在する可能性は0.94まで高まったと考えるんだ．ね，面白いだろ．詳しくはまた説明するよ．

Aさん，B君：楽しみにしてます！

Aさん：B先輩，元気でました？相関はむずかしいですよね．

B君：おかげさまで．そう言えばこないだ，ドクターの先輩からオマエいつもAさんと一緒につるんでるけど，あやしいなぁ．って言われたんだけど，それこそ，偽相関にもとづく因果関係のオーバーディスカッションだよね．

Aさん：B先輩，私は隣の研究室の先輩から，Bは鈍感なうえに変わらないから，大変だねって言われましたよ．先輩もどんどん中身を更新して変わってくださいね．なので今日のアイスはおあずけです．また明日！先輩！

参考文献

1) 川瀬雅也編著：生物学のための統計学入門，化学同人 (2009).
2) 川瀬雅也・松田史生：生物工学, **94**, 656 (2016).

（【第10回】は95巻10号に掲載予定です）