実践統計解析【第9回】

微妙な時のしきい値が肝心

松田　史生[1]*・川瀬　雅也[2]

C君：おはようございまーす．今日も頑張んなきゃ．

Aさん：……それって友達よりは上でも，お付き合いしているってほどではないってとこかなって思うんです．どうしたらいいんですかね，私．

B君：あ，C君いいところに来たね．今，Aさんの悩み事を聞いてるんだけど……

Aさん：ただの友達とお付き合いのしきい値って何で決まると思う？

C君：……え…？えええええええ？

B君：いやあ，むずかしいよね，C君．

C君：僕，忘れ物したので帰ります……でわぁ．

B君:あれ，C君えらい動揺したみたいだけど，何かあったのかな？それで，相談してきた部活の後輩には，なんてアドバイスしたの？

しきい値をどう設定するのか？

Aさん：微妙な時期が大事だから丁寧に頑張れっていっておきました．

B君：うまくいくといいねぇ．そういえば，今，データ解析に使っている相関係数も，微妙に相関があるときが厄介なんだよね…

Aさん：どういうことですか？

B君:いま，20条件で測定したメタボロームデータを使って，代謝物含量の増減パターンが似た代謝物のクラスターを抽出しようとしているんだよね．ロイシンとイソロイシン含量のピアソン相関係数は$r > 0.683$となって，しきい値0.65より大きいので，相関ありと見なす，という作業をしてるんだ（図1）．

＜リスト1＞

```
import matplotlib.pyplot as plt
from scipy.stats import pearsonr,spearmanr #相関係数を計算するモジュールをScipyからインポート
leu = [1.7,6.2,4.3,5.3,4.7,7.5,5.7,3.9,7.1,3.9,7.5,5.0,4.8,5.9,5.8,7.5,5.1,6.9,6.3,7.3] #ロイシンのデータ
ile = [3.4,11.7,12.1,13.1,13.3,13.6,13.6,14.3,15.0,9.5,16.7,17.0,17.3,17.7,18.6,19.1,19.2,21.2,21.5,21.9] #イソロイシンのデータ
r, pvalue = pearsonr(leu, ile) #ピアソン相関の計算
print("Pearson correlation r:", r)
r, pvalue = spearmanr(leu, ile) #スピアマン順位相関の計算
print("Spearman correlation r:", r,)
plt.scatter(leu, ile)
plt.show()
```

＜実行結果＞

```
Pearson correlation r:  0.683
Spearman correlation r: 0.529
```

B君：それをすべての代謝物の組合せで調べて，こういうような相関ネットワーク（図2）を書こうとしているんだよ．

Aさん：クラスターがわかりやすくて，かっこいいですね．これのどこが問題なんですか？

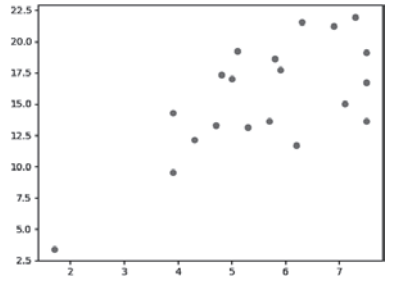

図1．20サンプル中のロイシン，イソロイシン含量の散布図．
横軸：ロイシン，縦軸：イソロイシン．

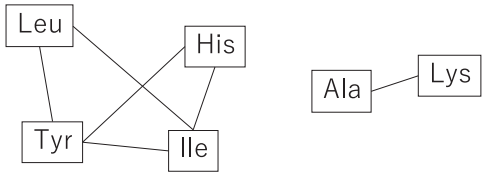

図2．相関ネットワークの例

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

B君：相関の有無を判定するには，しきい値がいるだろ？そのしきい値をいくつにしたらいいんだろ？というのが問題なんだわ．論文では，$|r| = 0.6, 0.65, 0.7$くらいが用いられているんだけどね．これってどういうふうに決まってるのかって知っている？

Aさん：知りません，でも確か，相関係数の検定法ってありませんでしたっけ？

B君：それはサンプル（データ）が正規分布に従う時だよね．でも今回はぜんぜん正規分布には見えないんだよね（図1）．

Aさん：ほんとだ．外れ値みたいなのもありますね．

B君：それで，外れ値の影響を受けにくい方法としてスピアマンの順位相関も試してみてるんだけど，r = 0.529になっちゃったんだよね．やっぱり，しきい値をどう設定したらいいのかわからないんだよなぁ．

Aさん：たしかに．これはX教授に聞きにいきましょうか．

相関係数の95％信頼区間を推定する

Aさん・B君：X教授．こんにちは．……というわけなんですが，どうするといいでしょうか？

X教授：ふむふむ．しきい値はいつもむずかしいなぁ．まずは，母数とサンプルの関係を復習しておこう．平均の場合，データ（標本集団）から計算した平均値μ（標本平均）が，母集団の平均（母平均）の推定値だ，っていうのはいいよね．それから，母集団が正規分布に従い，サンプル数が十分多い場合，標本標準偏差σを計算すれば，母平均の95％信頼区間はおおよそ$95\mu\text{-}1.96\sigma \sim \mu + 1.96\sigma$となる．

Aさん：統計の講義の最初のほうで習いましたね．

X教授：同じように，データ（標本集団）から計算した標本相関係数rにも95％信頼区間が存在する．もし，その95％信頼区間にゼロが含まれていなかったら，母相関係数はゼロじゃない可能性が高い．つまり相関があるといえる．

B君：でも，正規分布が仮定できないときって，どうやって95％信頼区間を推定すればいいんでしょうか？

X教授：データだけから95％信頼区間を推定する方法があるんだ．今回はブートストラップ法を勉強してみよう．

Aさん：前回の階層クラスター分析でも出てきましたよね．

X教授：たとえばB君のデータは，20組のロイシンとイソロイシンの含量のデータがある．これを使ってみよう．

①20組からランダムに一つ選ぶ．その後，選んだ組を戻してまた一つ選ぶ．この復元抽出を20回繰り返す．

②作成したデータで相関係数を計算する．

③①と②を可能な限り大量に，たとえば1万回とか繰り返す．

④1万個の相関係数を小さい順に並べる（ソート）．

⑤次に計算したこの数列の2.5と97.5パーセンタイル点（前から2.5％と97.5％番目の点，251番目と9750番目の点）の値が，95％信頼区間の下限と上限の推定値となる．ブートストラップ用のモジュールsklearn.utils.resampleを使ってやってみよう．

＜リスト2　ブートストラップ＞

```
from sklearn.utils import resample　#ブートストラップ用のモジュール
from scipy.stats import pearsonr,spearmanr
leu = [1.7,6.2, 省略6.3,7.3]
ile = [3.4,11.7, 省略21.5,21.9]
data = []
for i in range(10000):
    tleu, tile = resample (leu, ile,n_samples = 20) #leu, ileから20回復元抽出したリストを作成
    r, pvalue = spearmanr(tleu, tile)
    data.append(r)
data.sort()　#データをソート
print(data[251],data[9750]) #95%信頼区間の下限と上限を表示
```

実行結果（毎回微妙に異なるが類似の値になる）

0.208　0.824

Aさん：なんか騙されたみたいな気もしますね．ヨーロッパの古い民話で，ほら吹き男爵が，追っ手から逃れるべく，自分の“革靴のひも”（＝ブーツのストラップ）を引っ張って空中に飛び上がったという逸話があるそうです．手元のデータから統計量の推定を行うので，ブートストラップ法という名前がついたみたいですね．

B君：これで，正確に推定可能なんですか？

X教授：実用的には十分といわれている．でも，ものすごく正確な推定とは言えないため，⑤の手順を改良したさまざまな手法が提案されている[1)]．

Aさん：95％信頼区間の下限は0.208となっていて，95％信頼区間の上限と下限の間にゼロがないですね．これも相関がある根拠の一つになりますね．

B君：でも，今回の解析は，100種くらいの代謝物について総当たりで相関係数を計算するんです．そうすると，数千通りの組合せでブートストラップ法を行うと，ものすごく時間がかかってしまうことになりますよね．

ヌル分布を推定する

X教授：なので，次はヌル分布について考えてみよう．
B君：ヌル分布ってなんでしたっけ？
Aさん：まったく相関のないデータ間でも，計算すれば何かしら相関係数が出ますよね．この偶然に得られる相関係数を無限個集めたものがヌル分布ってことでよかったですかね？
X教授：そうだね．ヌル分布は検定の基礎になっているんだったよね．ヌル分布を推定してみようか．
①20組のロイシンとイソロイシンの含量のデータのうちイソロイシンだけランダムに順番を入れ替える．
②作成したデータで相関係数を計算する．
③①と②を可能な限り大量に，たとえば1万回とか繰り返す．
④1万個の相関係数を小さい順に並べる（ソート）．
⑤この数列の2.5と97.5パーセンタイル点の値が，ヌル分布の95 %信頼区間の下限と上限の推定値となる．
⑥もし，実測した相関係数がこの95 %信頼区間の外側にあれば，有意水準αが0.05で相関があるといえる．
Aさん：ブートストラップ法と似た方法で無相関なデータをいくつも作って，ヌル分布を作るんですか．

＜リスト3　ヌル分布の推定＞

```
from sklearn.utils import resample
from scipy.stats import pearsonr,spearmanr
leu = [1.7,6.2, 省略 6.3,7.3]
ile = [3.4,11.7, 省略 21.5,21.9]
data = []
for i in range(10000):
    tile = resample(ile,replace = False)# replace = Falseとすると，復元なしの抽出（ランダムな並べ替え）になる．
    r, pvalue = pearsonr(leu, tile)
    data.append(r)
data.sort()
print(data[250],data[9750])
```

実行結果（毎回微妙に異なるが類似の値になる）
-0.424 0.463

B君：お，出てきましたね，このヌル分布の95 %信頼区間の推定値が，しきい値の一つの目安．ということですか？でもこの方法でも代謝物の組合せごとにヌル分布を作り直さないといけないですね．
X教授：ここでスピアマンの順位相関の出番だ．
Aさん：なるほど．そういうことですか．スピアマンの順位相関は順位に変換したデータを用いるわけだから，どの代謝物の組合せでもヌル分布は同じ，になるはずですね．
X教授：その通り．つまりヌル分布の95 %信頼区間の下限，上限値はサンプルサイズごとに，事前に計算しできる．有意水準αが0.05でn = 20の時の上限は0.446みたいだね（表1）．
B君：じゃ，この値より大きければ，相関関係ありとみなせるってことですか？僕のデータを計算してみると，100代謝物の総当たりでスピアマンの順位相関の相関係数を調べると全体で650個の組合せに有意差ありってことで，一件落着ですね．

その関係はまぼろし？

X教授：いや，そうはいかないんだ．あくまでも95 %信頼区間の上限と下限値ってことは，無相関のデータでも5 %の確率で，有意差ありって出てしまうことになる．
Aさん：偽陽性でしたっけ？
B君：そんなの誤差みたいなもんじゃないの？ 100化合物の組合せだと100 *101/ 2 * 0.05 = 252.5. あちゃー．250個くらいは，偽陽性が得られると期待されるってことですね．ということは，有意差ありと判定したグループのうち，253/650 = 38.9 %くらいは偽陽性と推定できるだってことですね．
X教授：今計算した値が偽陽性率False discovery rate（FDR）というものだ．このFDRをきちんと把握できていないと，えらいめにあうよ．
B君：まぼろしの相関関係が4割も含まれている結果で

表1．スピアマンの順位相関のヌル分布の95 %信頼区間の上限値（100万回の試行結果）

	95 %信頼区間の上限値				
n	α 0.05	α 0.02	α 0.01	α 0.005	α 0.002
10	0.636	0.733	0.781	0.818	0.866
15	0.517	0.603	0.657	0.696	0.746
20	0.446	0.520	0.566	0.609	0.657
30	0.362	0.425	0.468	0.503	0.548
100	0.196	0.232	0.468	0.279	0.307

何か議論したら，それこそまぼろしの誤った結論に到達するかもしれないですね．いい夢見させてもらったよって，涙目になりかねない．

X教授：サンプル数と有意水準ごとに，スピアマンの順位相関のヌル分布の95 %信頼区間の上限値を計算したのが表1だ．

サンプルサイズが小さいとややこしい

X教授：この表を使って考える．まず，見つけたい相関のレベルを決める．たとえば，生物工学分野だと分子と分子の相互作用などを反映した直接的な関係を見つけたいので，$|r| > 0.6$くらいの相関を見つけたくなる．

Aさん：経済学だと，いろいろな要因が入り込むことが事前にわかっているので，$|r| > 0.4$くらいの相関を見つけることもあるみたいですね．

X教授:でだ．まず，議論したい相関のレベルを決める．次にそれを可能にするサンプルサイズを決める．もし，$|r| > 0.4$くらいの相関を見つけたいなら，サンプルサイズは10とか20とかではダメだろ？

B君：表1を見ると，最低でも30くらいは必要ですね．もし，$|r| > 0.6$の相関を考えるなら$n = 10$では全然ダメで，$n = 15$くらいでしょうか？

X教授：なので，$|r| > 0.6$くらいの相関を議論するための最低サンプル数は$n = 15$といわれる．まず，$n = 15$，できれば$n = 20$のデータを集めよう．B君のデータは合格だ．

B君：よかった….

X教授：ロイシンとイソロイシン間の相関1つだけを考えるときはこういう理路になる．B君のデータは，$n = 20$で標本相関係数$r = 0.529$だよね．標本相関係数が0.446（有意水準αが0.05，$n = 20$の時の上限値）より大きい．なので，相関係数はゼロである，という帰無仮説は有意水準αが0.05で棄却できる．ロイシンとイソロイシン間に統計的に優位な正の相関があるとは言える．次に前に説明した方法で，標本相関係数から母相関係数の95 %信頼区間を推定すると，0.108～0.814になる．

B君：ものすごく95 %信頼区間の幅が広いですね．しきい値の$|r| > 0.6$からすごくはみ出てる．

X教授:つまり，分子メカニズムから期待される，$|r| > 0.6$の相関が確実にあるのかどうかは判断できない．

Aさん：なるほど．となると結果の解釈は，「ロイシンとイソロイシン間には有意水準αが0.05で統計的に有意な正の相関があると言える．しかし，分子メカニズムから期待される$|r| > 0.6$の相関の有無については今回のデータからは判断できない．データ数を増やした再解析や，他の実験での検証が求められる」っていう感じですか？

X教授：次に，100個の代謝物の相関ネットワークを作るとする．$N = 20$のデータなので，有意水準αを0.005に設定し，表1の閾値である$|r| > 0.609$となった相関を選ぶと，その時の偽陽性の期待値は，100*101/2*0.02 = 25個程度になる．

Aさん：しきい値$|r| > 0.657$（有意水準α 0.002に相当）に設定すると，偽陽性の期待値は10個まで減りますね．

B君：100代謝物の総当たりでスピアマンの順位相関の相関係数を調べると（計算している），$|r| > 0.609$，$|r| > 0.657$となった組合せは，それぞれ150と100個になりました．偽陽性率はそれぞれおよそ17 %と10 %ですね．

Aさん：この場合，$|r| = 0.65$にしきい値を設定すると偽陽性率10 %の相関ネットワークを作ることができる，ということですね．

X教授：あと，サンプルサイズが小さいときに注意してほしいのは，しきい値を$|r| = 0.65$にする主な目的は，偽陽性の数をコントロールするためであって，母相関係数が$|r| > 0.65$となる相関を抽出するためじゃないということだ．上でふれたように，標本相関係数が$|r| > 0.65$であっても，母相関係数が$|r| > 0.65$になるとは限らないからね．

Aさん：なるほど，相関係数のしきい値を$|r| = 0.65$に設定しました，って言われると，母相関係数が$|r| > 0.65$の組合せを抽出したと見なしてしまいそうですが，サンプルサイズが小さいときは，気をつけなきゃいけないということすね．

B君：でも，しきい値を厳しくしすぎると，逆に「本当は相関があるのに誤って相関がない（偽陰性）」と判定してしまう場合も増えませんか？

X教授：もちろんその通りだ．今回の場合，かなり多数の偽陰性が出ていると思う．なので，FDRを確認し，許容できるレベルにできるもっとも甘いしきい値が一番妥当じゃないかな．表1を見るとサンプルサイズが$n = 20$のときにしきい値を$|r| = 0.65$に設定すると有意水準αを0.002にできるよね．$|r| = 0.65$程度のしきい値がよく採用されるのは，多くの場合，このあたりがちょうどよい妥協点になるからだと思うよ．

サンプルサイズが大きいときは本来の意味で

Aさん：教授，網羅的な遺伝子発現データで，サンプルサイズが100や1000を超えるような場合でも，しきい値として$|r| = 0.6$くらいが使われているようなのですが．

B君：さっきの議論だと，サンプルサイズが100だったら，しきい値は$|r| = 0.3$くらいが妥当，となりませんかね？

X教授：サンプルサイズが100や1000を超えてくると，標本相関係数がほぼ母相関係数に等しくなるはずなんだ．95％信頼区間も非常に狭くなる．それから，しきい値を$|r| = 0.6$に設定したら，誤って偽陽性になる可能性もほぼゼロになる．

Aさん：ということは，サンプルサイズが大きい場合のしきい値は，本来の意味で設定されているということですね．同じメカニズムで発現が制御される2遺伝子のmRNA量は$|r| > 0.6$くらいの相関を示すはずだ，という仮説が反映されている，のかな．

B君：たしかに，$r = 0.3$くらいの相関が統計的に有意だと言っても，生物学的にどういう意味があるのか？というのは議論したい生命現象によりけりですね．やはり$|r| > 0.6$くらいないとねぇ．

Aさん：しきい値，奥が深いですね……．これだけで4ページ半も語れてしまうとは…．

X教授：微妙なところなので，丁寧にがんばるしかないね．けど，なんだかんだ言って，どんな場合でも，しきい値を$|r| = 0.6 \sim 0.7$に設定するのが良さそうだっていうのは，ちょっとおもしろいね．これ以外のしきい値を設定するのはよほどの理由がいるよね．

データベース検索

B君：しきい値はデータベース検索にも出てきますね．たとえば，DNAやアミノ酸配列の相同性検索を行うときや，プロテオミクスでMS/MSスペクトルデータでペプチド同定するときにも，しきい値を設定します．

Aさん：最近勉強したんです．任せてください．BLASTアルゴリズムでは，2つの配列の相同性スコアの計算法が決まっています．それから，ランダムな配列同士が偶然示す相同性スコアのヌル分布が，数式で記述されています．なので，2つの配列の相同性スコアから，「2配列間に相関はない」という帰無仮説のp値も計算できる．という仕組みです．

X教授：となると，1000件と100万件の配列データがあるデータベースを検索したとき，偽陽性ヒットが1つ出ると期待できるp値はそれぞれ10^{-3}と10^{-6}となる．このようにBLAST検索でも大規模なデータベースを検索するときは，しきい値を厳しくする必要がある．

B君：プロテオミクスのペプチド同定には，ターゲット・デコイ法がどうのこうのって聞いたような．

X教授：MS/MSスペクトルデータを用いたペプチド同定でも，類似性スコアを計算する．でも類似性スコアのヌル分布を推定するいい方法がない．そこで，ある生物の全タンパクから作成したデータベース（ターゲット）と，全タンパクの配列を逆向きにした偽物のタンパクから作成したデータベース（デコイ）を用意する．あとは，たとえば，10,000個のMS/MSスペクトルでターゲットとデコイのデータベースそれぞれでペプチド同定を行い，ターゲットで1,000個，デコイで50個同定できたら偽陽性率は50/1000＝5％と考える，という考え方になる．

Aさん：いずれにせよ，しきい値を決めるには，微妙なところなので，丁寧にがんばるしかない，ってことですね．わたしも頑張んなきゃ．

参考文献

1) Derek, A. Roff 著，野間口眞太郎 訳：生物学のための計算統計学―最尤法，ブートストラップ，無作為化法―，共立出版 (2011).