実践統計解析【第11回】

p値とサンプルサイズ

松田　史生[1]*・川瀬　雅也[2]

Aさん：先輩！Natureに「“統計的に有意差なし”もうやめませんか？」という記事が出てたみたいなんですけど．どういうことなんですかね？[1)]

C君：僕もネットのニュースサイトでも見ました．

B君：じゃあ，C君のデータの解釈を相談しに，X教授のところに行ってみようか．

p値はあくまで基準の一つ

Aさん:あの，先日のNatureの記事についてなんですが，実際私たちも困っているんです．

X教授：最近特にあの記事が注目されているけど，以前JBBにも類似の指摘が出てたくらいの長年の懸案なんだよ[2)]．

Aさん：最近，C君が物質Pよりも強い細菌の増殖抑制作用を示す物質のスクリーニングを行ったところ，新しく候補になりそうな物質Qが見つかりました．増殖抑制率のデータです（表1）．

表1．増殖抑制率

物質P	新規物質Q
47	46
48	55
49	54
47	60
46	45

C君：そこでスチューデントとウェルチのt検定（両側）をしてみました．

```
import numpy
from scipy import stats
Pdata= [47,48,49,47,46]
Qdata = [46,55,54,60,45]
t, pvalue = stats.ttest_ind(Pdata, Qdata)
print("Student's ttest pvalue:",pvalue)
t, pvalue = stats.ttest_ind(Pdata, Qdata, equal_var=False)
print("Welch's ttest pvalue:",pvalue)
結果
Student's ttest pvalue: 0.1502866740442524
Welch's ttest pvalue: 0.18255295912949338
```

C君：ということで，p値が0.05より大きいので，有意差なし，この新規物質Qには効果がない，という結論になりました．やっぱり，この物質はあきらめて，他をもっと探す方がいいでしょうか．

X教授：ふむふむ．それで？

B君：有意差検定で有意差が出ないなら効果がないんじゃないの？

Aさん：統計処理の結果を見ると，そういうことになるけど，何か納得できない．

C君：基本に戻って手法を確かめたのですが，

帰無仮説 H0：増殖抑制率に差はない
対立仮説 H1：増殖抑制率に差がある
有意水準 α 0.05

を設定し，p値を計算する．p値が有意水準αより小さい場合は,帰無仮説を棄却して対立仮説を採用する．つまり，増殖抑制率に差がある，と結論付けると講義で習いました．

X教授：では，p値が有意水準αより大きい場合はどうするって習った？

B君：そりゃ，帰無仮説が棄却できないんだから，帰無仮説を採用して「増殖抑制率に差はない」と考えるんじゃないの？

Aさん：いや，先生は講義でそうは言っていなかったような気がするんですが．それから，Natureの記事も統計的に有意差なしの解釈に問題があることを指摘する記事でした．

X教授：その通り．この帰無仮説と有意水準αを使った検定手法は，フィッシャーが1935年に行った偉大な発明だ．まず，有意差があることを「直接示す方法がない」ということが問題だった．そこで，差がないという帰無仮説を棄却するという間接的な方法を発見したことがすごい点だ．帰無仮説が成り立つとすると，観察されたデータはめったに起きない（有意水準α以下の確率でしか起きない）はずだから，帰無仮説は棄却できるとした．

一方，この方法の欠点は，結果の解釈が難しいことだ．特に，p値が有意水準αより大きい場合の解釈を間違えやすい．「観察されたデータでは，帰無仮説を棄却できない」というのが正解に近いと思う．

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：ややこしすぎるのですが…
Aさん：うーんと，実際のデータで確認できるのは，「帰無仮説が棄却できるかどうか」である．棄却できた時は，対立仮説を採用する．
C君：そこで，棄却できなかった時は，「実際のデータは，帰無仮説を棄却する証拠としては不十分だった」と考えるんですね．
X教授：さらに，帰無仮説が棄却できなかったからといって，帰無仮説を採用する根拠にはならないわけだ．
B君：となると，帰無仮説を採用も，棄却もしない．増殖抑制率に差があるか，ないか判断できない．ということですか？これまた中途半端ですね．
X教授：有意差検定の結果の解釈が難しいのは，帰無仮説を棄却できなかった時には，「結論が出ない」点にある．要するに，有意差検定で「有意差がない」と結論付けることは原理的に不可能なんだ．が，最近の研究の半分くらいは，ここを間違えて「有意差がない」としていたという報告だ．
B君：でも，それの何が問題なんですか？
C君：今回の例でいうと，「新規物質Qと物質Pの増殖抑制率に差はない」と結論付けると何がまずいかですね．実際は，新規物質Qに強い活性があった場合，重要な知見を見逃してしまうことになります．
B君：それは，C君の実験がへたっぴなのがよくないんじゃないの？
X教授：Natureの記事の焦点はそこなんだよ．
Aさん：なるほど，データの取得方法に問題があって，有意差を検出できないのだったら，「有意差がない」と結論づけるのは誤りですよね．

対策1：ばらつきを小さくする

C君：じゃあ，どうすればいいでしょうか？

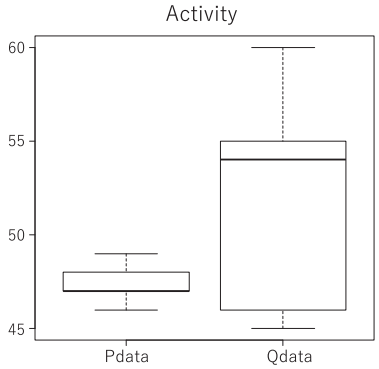

図1．C君のデータの箱ひげ図

Aさん：こういう時はばらつきをしらべるんでしたよね．箱ひげ図を書いてみましょ（図1）．

＜リスト1＞

```
import matplotlib.pyplot as plt
Pdata= [47,48,49,47,46]
Qdata = [46,55,54,60,45]
fig, ax = plt.subplots()
ax.set_title('Activity')
ax.set_xticklabels(['Pdata', 'Qdata'])
ax.boxplot((Pdata, Qdata))
plt.show()
```

B君：物質Qは作用が安定していないから，ダメなんじゃない？ばらつきが大きすぎるよ．
Aさん：でも，平均値は物質Qのほうが大きいんですよね．データのばらつきが大きいんですけど．
C君：それでAさんと相談して，結果がばらつく原因を考えてみました．

1. 物質Pは試薬として購入したが，新規の物質Qは植物から抽出したので不純物が混じっている可能性がある．
2. 物質Pは水に溶解して培地に添加した．新規物質Qは水に難溶だったので，エタノールに溶解して添加した．
3. 実験は，前半（3回）と後半（2回）を別の日に行った．
4. 実験後半は抽出しなおした新規物質Qを使った．

Aさん：となると，極力精製した新規物質Qを使って，1回の実験で，物質Pもエタノールにそろえて実験をしてみるのはいいかもしれませんね．
X教授：実験のばらつきをまず減らさないとね．以前，『生物工学会誌』94-8号と95-2号で取り上げているので参考にしてほしい．また，Natureの記事では，薬の効き目について，今見たのと同じことが起こっていることを例に出している．そして，有意差がないから“差がない”としている論文が多数存在することも示して，そろそろ有意差という考え方から卒業しよう“Retire statistical significance”と呼び掛けている．そして，信頼区間を利用しようとしている．

対策2：反復数を増やす

Aさん：その次は，実験の反復数（サンプルサイズ）を

増やすんでしたよね.

B君：でもなんで，サンプルサイズを大きくするのが有効なの？

C君：t検定の帰無仮説を正確に言うと「2群の平均値に差がない」になります．サンプルから推定する平均値の信頼区間（ばらつき）はサンプルサイズの平方根に比例して狭くなっていきます．なので，サンプルサイズを増やす＝ばらつきが小さくなる＝2群の平均値の差を検出しやすくなる，はずです.

B君：じゃあさあ，たとえばさっきのn ＝ 5のデータを2つつないでn ＝ 10にしてみたりすると，結果がかわるってこと？

＜リスト2＞

```
Pdata= [47,48,49,47,46,47,48,49,47,46]
Qdata = [46,55,54,60,45,46,55,54,60,45]
t, pvalue = stats.ttest_ind(Pdata, Qdata)
print("Student's ttest pvalue:",pvalue)
t, pvalue = stats.ttest_ind(Pdata, Qdata, equal_var=False)
print("Welch's ttest pvalue:",pvalue)
結果
Student's ttest pvalue: 0.028201668590677034
Welch's ttest pvalue: 0.03922221351324858
```

Aさん：2種類の検定方法でp値が有意水準α0.05より小さくなりましたね．ということは…

C君：この結果はあくまで，仮想的なものですけど，今の実験のばらつきのままでも，サンプルサイズを2倍にすれば新規物質Qの有効性を実証できるかもしれない！ということでしょうか？

対策3：効果量を使う

X教授：すこし展望が開けてきたかな？

C君：はい．急いでラボに戻って実験やり直してみます.

Aさん：C君頑張ってね．あとで見に行くし.

B君：でも，今回はたまたま2倍したらうまくいったけど，事前にサンプルサイズを決めることはできないんでしょうか？

X教授：できる．というより，やるべきだと思うんだけどね．Aさん，平均値の差の大きさと，実験のばらつき（標準偏差）とサンプルサイズがp値に与える影響を考えてみて.

Aさん：表2でいいでしょうか？有意差が検出しやすい（p値が小さくなる）のは，平均値の差が大きく，実験のばらつきが小さく，サンプルサイズが大きいときですよね.

X教授：このうち，「平均値の差」と「実験のばらつき（標準偏差）」に注目したのがコーエンの効果量という概念だ．効果量は，平均値の差が大きく，実験のばらつきが小さいときに大きくなる性質がある．下の表2にも効果量を記入しておいた.

表2．p値との関係

小	p値	大
大	平均値の差	小
小	実験のばらつき	大
大	効果量	小
大	サンプルサイズ	小

B君：C君がいないので，効果量について，いまネットで調べました．P，Q 2群のサンプルサイズが同じ時の効果量dは

d = (Qの平均 − Pの平均) / sqrt((Qの標準偏差^2 + Pの標準偏差^2)/2)

で計算できます．もし，P，Qの標準偏差が同じだったとすると

d = (Qの平均 − Pの平均) / 標準偏差

となりますから，実験のばらつきに対する，平均値の差の大きさの指標ですね.

Aさん：実験のばらつきに対する効果の大きさ，という効果量ですね．じゃあ，早速C君のデータで計算してみましょう.

＜リスト3＞

```
import numpy
Pdata= [47,48,49,47,46]
Qdata = [46,55,54,60,45]
d1 = numpy.mean(Qdata)- numpy.mean(Pdata)
d2 = numpy.sqrt((numpy.std(Qdata)**2+
numpy.std(Pdata)**2)/2)
d = d1/d2
print("effect size:",d)
結果
effect size: 1.1249667725023929
```

Aさん：効果量d = 1.12って出ましたけど.

B君：今調べたところ，コーエンの効果量には正負があります．正になったのは，Qの平均がPより大きいからです．効果量の指標として，小 $|d| > 0.2$，中 $|d| > 0.5$，

大 $|d| > 0.8$ が提案されていますね.

Aさん:ということは,C君のデータでも新規物質Qには,物質Pに比べて十分大きな効果があるといえますね.

X教授:ときどき,p値に加えて,効果量も論文で報告しようという提案があるんだけど,あまり広く受け入れられてはいないかな.今回の結果は,私だったら「有意水準α 0.05のWelch検定では有意と認められなかったが,コーエンの効果量 d は1.12と大きかったことから,新規物質Qの阻害効果は物質Pと比べて高い可能性がある.さらに,サンプルサイズを増やすなどの検討を加えることで,より明確な結果が得られると期待される.」くらいにしておくかな.

対策4:サンプルサイズを推定する

B君:じゃあ,サンプルサイズを増やせばいい,ってことでしょうか?たとえば,もし可能だったらn = 1000とか.

X教授:でも,むやみに大きくしても実験が大変なだけだよね.必要最低限なサンプルサイズを把握することが重要だ.有意水準αが0.05というのはどういうことかというと,本当は平均値に差がないのに帰無仮説を棄却してしまう誤り(第一種の過誤)が20回に1回は起きるということを示している.一方,本当は平均値に差があるのに,帰無仮説を棄却できない誤り(第二種の過誤)が起きる確率をβとする.さらに(1-β)を検定力(power)と呼び,0.8以上になると良いとされる.で,サンプルサイズを大きくすると

＝＞第一種の過誤が起きる確率は減る.

＝＞第二種の過誤が起きる確率も減る.

という関係がある.当たり前だね.重要なのは,効果量,α,(1-β)の3つが決まると,サンプルサイズが決定するという性質がある点だ.

Aさん:たいていはα = 0.05,(1-β) = 0.8に設定しますから,要するに,効果量が決まると,必要なサンプルサイズも決定できるというわけですね.

X教授:pythonにはstatsmodelsというパッケージがあり,この計算ができる.そこに,もしpythonをAnacondaとしてインストールしていれば,デフォルトで使える.Aさん調べてみて.

Aさん:“python 検定力 サンプルサイズ statsmodels”で検索すると詳しいページがたくさん出てきました.tt_ind_solve_powerに効果量(effect size),α(alpha) 0.05,(1-β)(power) 0.8をそれぞれ入力すればいいようですね.

＜リスト4＞

```
#必要なモジュールのインポート
from statsmodels.stats.power import tt_ind_solve_power
import numpy
Pdata= [47,48,49,47,46]
Qdata = [46,55,54,60,45]
#効果量の計算
d1 = numpy.mean(Qdata)- numpy.mean(Pdata)
d2 = numpy.sqrt((numpy.std(Qdata)**2+
numpy.std(Pdata)**2)/2)
d = d1/d2
print("effect size:",d)
#サンプルサイズの推定
n = tt_ind_solve_power(effect_size=d, alpha=0.05,
power=0.8)
print("sample size:",n)
結果
effect size: 1.1249667725023929
sample size: 13.435468718120164
```

B君:C君の実験で有意差検定を行うための適切なサンプルサイズはn = 13くらいいってことか.N = 5では全然不足していたわけね.

Aさん:ちなみに,この方法を使うと実験のデザインをシミュレーションできます.A = 0.05,(1-β) = 0.8に固定して,効果量 d を3にすると,適切なサンプルサイズがn = 3になります.ということは,サンプルサイズがn = 3のときは,

・実験の標準偏差が3のときのt検定とは,A群とB群の平均値の差が,9以上があるかどうかを検定しようとしている.

・平均値の差が9の2群の差を検出するには,実験の標準偏差を3以下に小さくする必要がある.

などがいえるってことですね.

B君:よく分かりました.統計処理の結果は,あくまでも,議論の補強材料と,サンプルサイズを意識することが大事ということですね.

X教授:このことは,非常に大事なことなので,頭のどこかに置いてほしいと思うんだ.

文　献

1) Amrhein, V. *et al.*: *Nature*, **567**, 305 (2019).
2) Kawase, M. *et al.*: *J. Biosci. Bioeng.*, **100**, 116 (2005).