Series

正しい知識が捏造を防ぐ

データを正確に解釈するための6つのポイント

No. 5

大規模データの解析における問題点

DNAマイクロアレイによる遺伝子発現量の測定を例として

山田陸裕・上田泰己

はじめに

われわれは日々の研究活動において実験を行ない，その結果が自説を支持するものなのかどうかを客観的に判断するため，統計的な手法を用いている．しかし，DNAマイクロアレイに象徴されるような大量データを生産する実験技術の急速な進歩により，これまで，おもてだって問題になることの少なかった大規模データの解析に特有の問題が現われてきた．今後，われわれ生命科学者が大規模データを自分で解析したり，あるいは，他人の解析結果を解釈する必要にせまられたりする機会はますます増えていくだろう．本稿では，手はじめに統計的手法の基本的な考え方を再訪したのち，大規模データの統計的な取り扱いにおいて遭遇する多重検定の問題と，現在，用いられている代表的な対処方法を紹介する．

統計学再訪：仮説検定

①ある仮説のもとで，②データが得られる確率を考えることにより，③そのデータに意味があるのかどうかを判断する．この3ステップが本稿で扱う統計的な手法の考え方の根本である．以降，この考え方を具体的に使いながらデータを解析する．まず最初に，ひとつ例をあげて用語を導入しておく．

［例1］全部で11の実験を行なった．1～10番目は同じ条件，11番目は異なる条件での実験である．得られたデータのうち，1～10番目のデータと11番目のデータに差があるように思える場合，11番目のデータには客観的にも差があるといえるのか？　表1に，データの値を示す．

この問いに答えるため，さきの3つのステップをあてはめ，①“11番目のデータは1～10番目のデータと大差ない”とする仮説のもとで，②実際に11番目に得られたデータ以上に大きな差が得られる確率（これを*p*-value，*p*値という）を計算することにより，③11番目のデータと1～10番目のデータとのあいだに差があるのかどうかを判断する．もし，*p*値が十分に低ければ仮説が正しい可能性は低いのだから，10番目までのデータと11番目のデータには差があると判定し，これらのあいだに“有意（significant）な差がある”という．判定の基準となる*p*値を有意水準という．また，“差がある”ことをいうために“差がない”という仮説をたてたが，このように否定されることを意図してたてられる仮説を帰無仮説（null hypothesis）という*1．

ところで，“大差ない”といういい方はあいまいなので，仮説を“1～10番目のデータと11番目のデータが同じ分布から得られた”といい換え，分布というものについてより具体的に考えてみよう．実験によって得られるデータは，毎回，ある値のまわりに分布するが，いくつかの代表的な分布パターンについては数学的な表現が可能である．なかでも，頻繁に用いられるのが正規分布（normal distribution）

Rikuhiro G. Yamada, Hiroki R. Ueda
理化学研究所発生再生科学総合研究センター システムバイオロジー研究チーム
E-mail：rikuhiro@cdb.riken.jp，uedah-tky@umin.ac.jp
URL：http://www.cdb.riken.jp/lsb/jpn/index.html

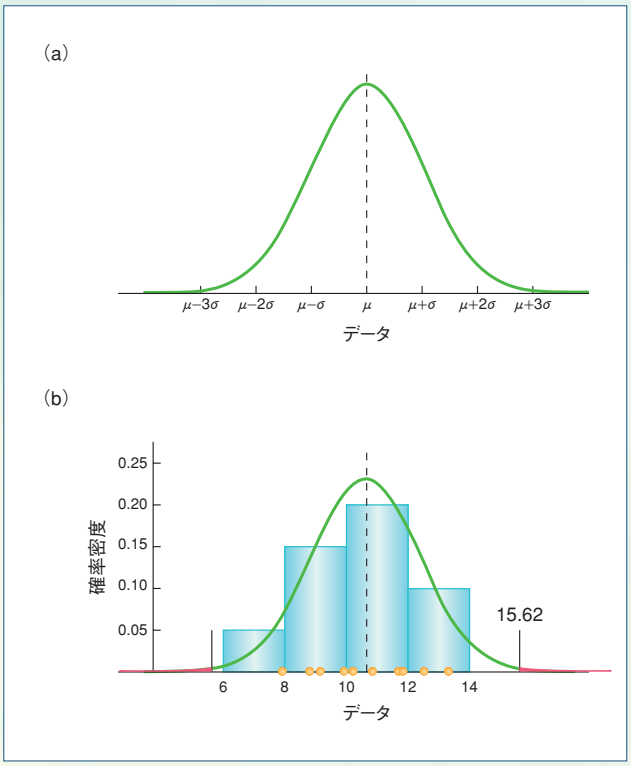


図1　正規分布の例

(a) 平均 (μ) と標準偏差 (σ) の正規分布 $N(\mu, \sigma)$.
(b) 正規分布 N (10.6, 1.72) と，1～10番目のデータのヒストグラム．全体の面積が1になるようスケールをあわせている．黄色の点はデータ点．11番目のデータ (＝15.62) とその外側の領域の面積を赤色で塗っている (面積は，0.0038)．

とよばれる分布である*2．正規分布は，平均 (μ) と標準偏差 (σ) (もしくは，その2乗として求まる分散) をパラメーターとしてとり，$N(\mu, \sigma)$ と表記される (図1a)．たとえば，表1の1～10番目のデータの平均は10.6，標準偏差は1.72 (分散は2.97) と計算されるが，これらをパラメーターとしてとる正規分布は N (10.6, 1.72) と書くことができる*3．ここでは，表1の1～10番目のデータは，この N (10.6, 1.72) にしたがっていると考えることにする．

表1　[例1] の実験データ

実験	データ
1	7.93
2	11.71
3	8.81
4	11.83
5	10.83
6	9.93
7	9.16
8	12.49
9	10.21
10	13.32
11	15.62

11番目のデータ (=15.62) は，この N (10.6, 1.72) という分布の平均から5.02 (=15.62−10.6) だけ離れているが，これより離れた値が同じ分布から得られる確率 (すなわち，p 値) はどれくらいなのだろうか？　実は，正規分布は単に平均値のまわりにデータが多く分布することを表わすだけでなく，曲線と x 軸とのあいだの面積が対応するデータの得られる確率に等しくなるよう定義されている*4,*5．このため，分布の平均値 (=10.6) から5.02より外側の領域で曲線と x 軸とのあいだの面積を計算すれば11番目の値の p 値が得られる*6 (図1b)．実際にこれを計算してみると，p 値は0.01よりも小さい，すなわち，100回に1回もない事象であることがわかる．

以上を，さきの3つのステップにそってまとめると，①仮説“1～10番目のデータと11番目のデータが同じ分布から得られた”が正しいとすると，②11番目のデータが得られる確率は100回に1回もない，ということである．しかし，そんなに得られる確率の低いデータが現実に手元にあるわけだから，前提とした仮説が間違っているのだと考え，③“1～10番目のデータと11番目のデータは同じ分布から得

---

*1　**帰無仮説**：帰無とはずいぶん変わった響きの言葉を使うものだと思うが，否定されて無に帰することが意図されている，という程度で，あまり深い意味はない．また，帰無仮説を否定することを，棄却するという．帰無仮説が棄却されるときに採択される仮説を，対立仮説 (alternative hypothesis) とよぶ．これらは，統計検定では頻出する独特のいいまわしである．

*2　**正規分布**：ガウス分布 (Gaussian distribution) ともよばれる．

*3　平均は $\mu=\frac{1}{N}\sum_{i=1}^{N}D_i$，分散は $\sigma^2=\frac{1}{N-1}\sum_{i=1}^{N}(D_i-\mu)^2$ により計算する (D_i は i 番目のデータ，N はデータの数)．標準偏差は，分散の平方根である．

*4　正規分布を表現する関数 $f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$ は，一見，複雑なかたちをしているが，exp (　) のまえにある部分は範囲 ($-\infty, +\infty$) で積分したとき1になるようにするための係数で，exp (　) のなかは x と平均 (μ) との差が標準偏差 (σ) に比して大きくなるほど確率が小さくなることを表わしている．

*5　このような性質をもった関数は，確率密度関数とよばれる．統計学ではさまざまな分布とそれに対応する確率密度関数が準備されていて，正規分布はそのうちのひとつである．

*6　実際の計算は統計用のソフトウェアを用いれば容易に実行できる．また，統計学の一般的な教科書には正規分布の面積に関する一覧表がついているので，それを利用することもできる．ただし，多くの場合，N (0, 1) についての表になっているので，式 (D-μ) / σ (D：評価しようとしている値，μ：分布の平均値，σ：分布の標準偏差) により，適宜，スケールをあわせて一覧表と照らし合わせる．ここでの例でいえば，11番目のデータ (D = 15.62) から平均値10.6を引いて，さらに標準偏差の1.72で割った値 (＝2.92) により一覧表を利用する．

られたものではない”と判断する．

ここでは，p値が0.01という有意水準において判断を行なった．しかし，この0.01という値に明確な根拠はなく，慣習として0.01や0.05が有意水準としてよく使われているので，ここでもそれを踏襲したにすぎない．

ここでは，ひとつのデータについて，ほかの複数のデータと比較して差があるかどうかを判断した．現実的には，くり返し実験を行なって得た複数のデータについて，差があるかどうかを判断したい場合が多いだろう．以下では，そのような例について考える．

平均の差の検定：t検定

実験データの解析では，条件を変えて得られた2つのデータ群のあいだに差があるかどうかを調べたいことがある．

[例2] ある転写因子（遺伝子Aにコードされる）が，注目している遺伝子（遺伝子B）の発現を活性化するかどうかを調べたい．遺伝子Aを過剰発現させて遺伝子Bの発現量を計測する実験を処置実験とし，ダミーとなる遺伝子を過剰発現させて遺伝子Bの発現量を計測する実験を対照実験とする．これらの処置実験と対照実験をそれぞれ複数回行な

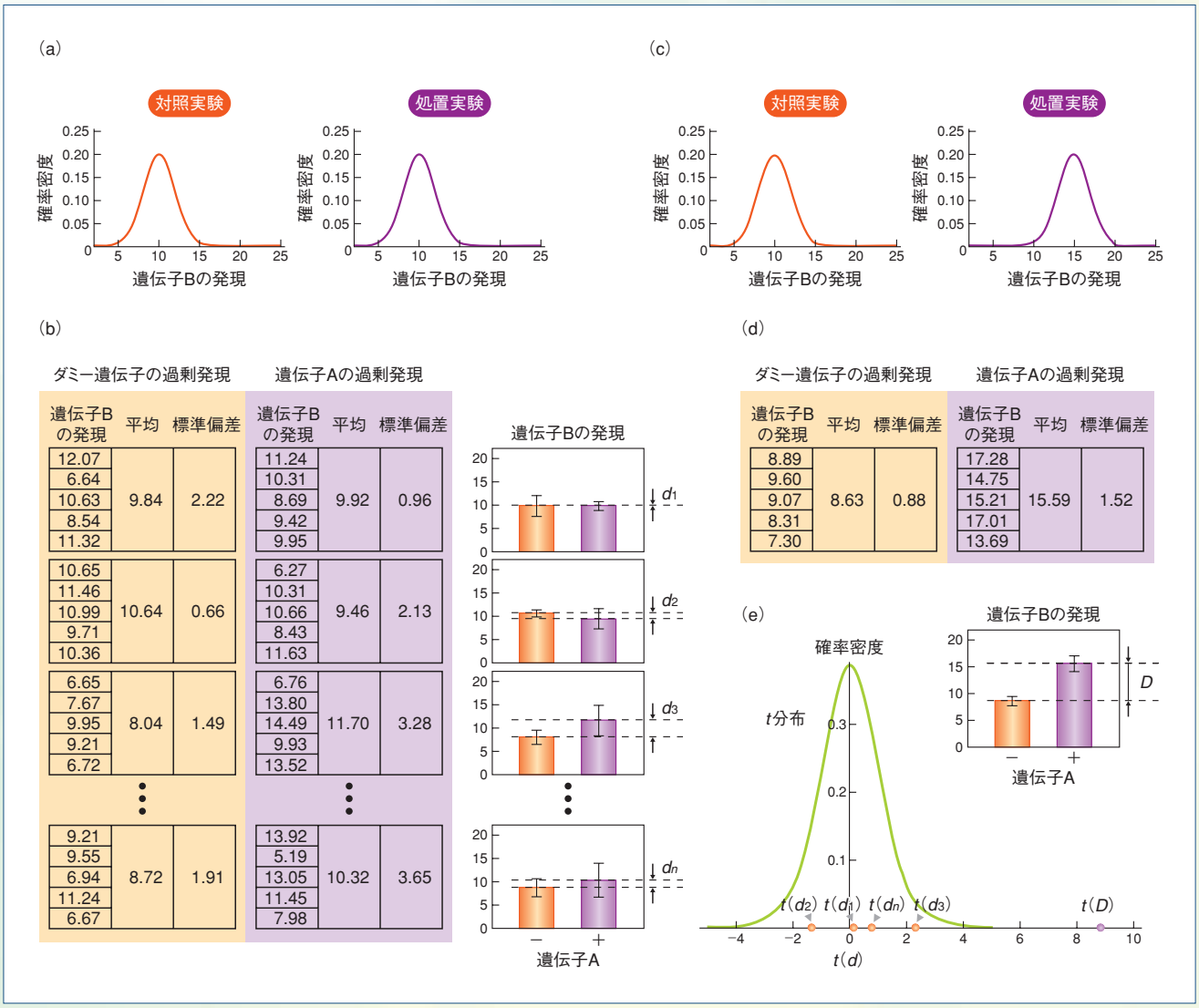
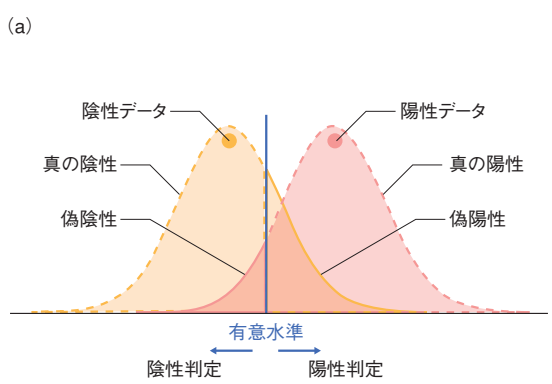


図2　t検定の例

(a)“2つのデータ群（対照実験と処置実験）に属するデータが平均の等しい2つの正規分布から得られたものである”という仮説にもとづく正規分布と，(b)実データの例．ここでは，対照実験と処置実験においてそれぞれ5回の実験を行ない，その平均値を比較している．仮説より両者の平均値には差がないので，平均の差（d_1，d_2，d_3，…，d_n）の分布は0を平均とするt分布(e)となる．(c)2つのデータ群（対照実験と処置実験）に属するデータが異なる正規分布から得られている場合の正規分布と，(d)実データの例．両者の平均値の差Dの得られる確率の低いことが，Dが仮説にもとづくt分布から大きく外れている(e)ことから判断できる．



(b)

実際＼判定	陰性	陽性
陰性	真の陰性	偽陽性
陽性	偽陰性	真の陽性

図3　有意水準と判定の種類

(a) 一部に重なりのある陰性データと陽性データの分布を，ある有意水準で陽性判定と陰性判定に分けたときの，各データのよばれ方．(b) これを表形式にまとめたもの

い，それぞれのデータ群の平均値に差があるかどうかを判定する．

このようなときも，冒頭で述べた，①ある仮説のもとで，②データが得られる確率を考えることにより，③そのデータに意味があるのかどうかを判断する，という考え方は同じで，ここで用いる仮説は，“2つのデータ群（対照実験と処置実験）に属するデータが，それぞれ平均の等しい2つの正規分布から得られたものである”というものである（ここでは，等分散性は仮定しない*7）．この仮説のもとでは，対照実験と処置実験の実験セットをくり返したときに得られる2つのデータ群の平均の差は，0を平均として分布することになる．そして，この仮説のもとで手元にあるデータが得られる確率を考える方法が，t検定とよばれる統計検定手法である（図2）．

t検定で注意が必要なのは，用いる仮説には，平均が等しいこと，データが正規分布から得られているという，2つの仮定が含まれていることである．もし，データが正規分布にしたがっていない場合は，仮に有意水準をクリアする結果が得られたとしても，それが2つのデータセットの平均の差によるものなのか，データセットが正規分布から外れていることによるものなのか，区別がつかない．このため，t検定を行なうまえに，データが正規分布にしたがっているとみなしてさしつかえないかどうかを可能なかぎりしっかりと確認しておくことが重要になる*7．しかし，現実的には，実験上の制約から正規分布であることを確認するのに十分なサンプル数を確保できない場合も少なくない．そのような場合には，とくに，t検定の結果で有意水準をクリアしていたとしても安心せず，ほかの手法による確認をあわせて行なうなどの配慮が欠かせない*8．また，散布図やヒストグラム（図1）などを描画してデータの分布を視覚的に確認することはたいへん重要である．実データの分布の仕方は実験によって千差万別だが，統計検定はそのなかでも代表的で数学的に取り扱える理想的なものを対象としていることをつねに念頭においておく必要がある．

ところで，t検定でも有意水準として0.05や0.01が一般的に採用されるが，どのような有意水準を選ぶにしても“仮説を前提とした場合には滅多に起こらない”ということを根拠に判定しているので，たまに起こる偶然によって間違った判定がなされてしまうことがありうる．のちに述べる多重検定では，統計検定を多数回くり返すためとくにこのことが問題となるので，ここで，判断の正解と間違いについて整理しておきたい．間違いは2種類に分けられ，ひとつは実際には差がない（陰性，negative）のに差がある（陽性，positive）と判断されてしまう偽陽性（false positive）とよばれる間違いであり，もうひとつはその逆に，実際には差があるのに差がないと判断される間違いで偽陰性（false negative）とよばれるものである（図3）．

また，図3からは，偽陰性を減らすためには有意水準を甘くすればよいが，そうすると偽陽性が増えてしまうことも理解できるだろう．つまり，有意水準を甘くすることで感度は上がっていくけれども，特異度が下がってしまうというわけである．ここにはトレードオフの関係があって，感

＊7　多くの解説書では，概念の解説を簡潔にするため等分散性も仮定するt検定が説明されることが多いが，より実用的なのは，等分散性を仮定しないWelch（ウェルチ）のt検定とよばれる手法であり，ここでもそれを想定している．また，正規性の確認にはコルモゴロフ-スミルノフ検定やシャピロ-ウィルク検定を用いることが一般的である．

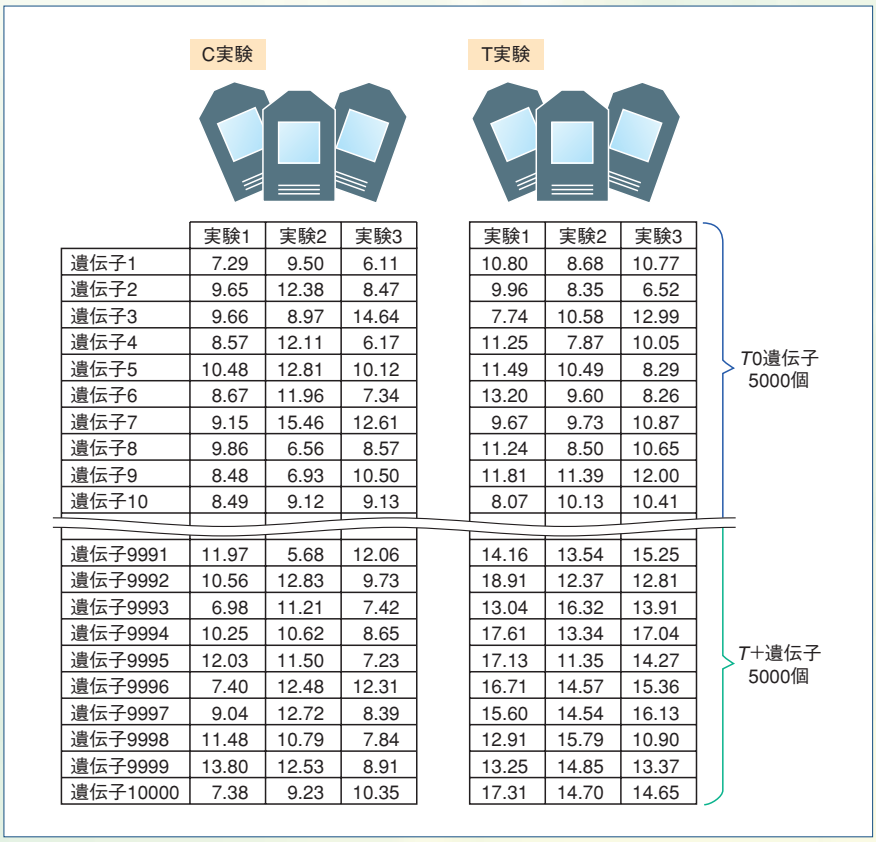
＊8　分布を仮定しないt検定に対応する検定方法として，マン-ホイットニーのU検定とよばれる手法がある．また，一般的に，分布を仮定しない検定方法は“ノンパラメトリックな検定方法”とよばれる．ただし，このノンパラメトリックな手法は一般に検出力に劣ると考えられる．

度と特異度，どちらか一方をたてればもう一方がたたない．ちなみに，感度（sensitivity）とは陽性データのなかで判定も陽性となるものの割合，特異度（specificity）とは陰性データのなかで判定も陰性となるものの割合のことである*9．

多重検定：有意水準の調整

ここまで，統計検定の考え方について再確認した．基本は冒頭で述べた，①ある仮説のもとで，②データが得られる確率を考えることにより，③そのデータに意味があるのかどうかを判断する，ということであった．そして，たとえば，*t*検定という統計検定法で有意であるという判断をした場合，それは“2つのデータ群が平均の等しい2つの正規分布から得られたものである”という仮説のもとで，手元にある2つの実験データ群の平均値の差が得られる確率が有意水準（たとえば，0.01以下）をクリアしていることを意味した．つまり，“この仮説が正しいとしたら，現に手元にあるデータが偶然に得られる確率は100回に1回もなく，したがって，仮説は間違っている可能性が高く，2つのデータ群の平均値には差がある”と判断したのであった．いちどの解析で扱う検定の回数が1～2回であれば，この0.01や0.05という有意水準は妥当なものであろう．しかし，いちどに多数回の検定を行なう場合はどうだろうか？

たとえば，異なる2つの条件でDNAマイクロアレイを使ったゲノムワイドな遺伝子発現量の測定を行なって発現量が有意に異なる遺伝子を探し出そうとする実験の場合，適当な前処理ののち*10，DNAマイクロアレイで測定した遺伝子の数だけ統計検定を行なうことになる．偶然だけでは100回中1回しか得られないような実験データでも，DNAマイクロアレイで測定できる遺伝子が10,000個あるとして，10,000回検定すれば100回も得られる．つまり，検定回数が数回の場合と多数回の場合で同じ有意水準を適用したのでは偽陽性が多くなりすぎることが憂慮される．では，どのようにして有意水準を調整すればよいのだろうか？



	実験1	実験2	実験3	実験1	実験2	実験3	
遺伝子1	7.29	9.50	6.11	10.80	8.68	10.77	
遺伝子2	9.65	12.38	8.47	9.96	8.35	6.52	
遺伝子3	9.66	8.97	14.64	7.74	10.58	12.99	
遺伝子4	8.57	12.11	6.17	11.25	7.87	10.05	
遺伝子5	10.48	12.81	10.12	11.49	10.49	8.29	*T*0遺伝子 5000個
遺伝子6	8.67	11.96	7.34	13.20	9.60	8.26	
遺伝子7	9.15	15.46	12.61	9.67	9.73	10.87	
遺伝子8	9.86	6.56	8.57	11.24	8.50	10.65	
遺伝子9	8.48	6.93	10.50	11.81	11.39	12.00	
遺伝子10	8.49	9.12	9.13	8.07	10.13	10.41	
遺伝子9991	11.97	5.68	12.06	14.16	13.54	15.25	
遺伝子9992	10.56	12.83	9.73	18.91	12.37	12.81	
遺伝子9993	6.98	11.21	7.42	13.04	16.32	13.91	
遺伝子9994	10.25	10.62	8.65	17.61	13.34	17.04	
遺伝子9995	12.03	11.50	7.23	17.13	11.35	14.27	*T*+遺伝子 5000個
遺伝子9996	7.40	12.48	12.31	16.71	14.57	15.36	
遺伝子9997	9.04	12.72	8.39	15.60	14.54	16.13	
遺伝子9998	11.48	10.79	7.84	12.91	15.79	10.90	
遺伝子9999	13.80	12.53	8.91	13.25	14.85	13.37	
遺伝子10000	7.38	9.23	10.35	17.31	14.70	14.65	

図4　DNAマイクロアレイ実験の例

＊9　感度は，真の陽性/（偽陰性＋真の陽性）．特異度は，真の陰性/（真の陰性＋偽陽性）．
＊10　たとえば，遺伝子の発現量は対数正規分布に近いと考えられているので，統計解析を行なうまえに発現量のlog値を計算するなどの処理がなされる．

Bonferroni 補正

ひとつの方法は，全検定中でひとつでも偽陽性が得られる確率がある目標値になるよう，各検定の有意水準を調整する方法である．たとえば，10,000回の検定を行なったとき偽陽性がひとつ得られる確率が0.01になるように，各検定の有意水準を0.01/10000とする．この方法はBonferroni（ボンフェローニ）補正とよばれる．シンプルだが，とくに検定回数が多いときには有意水準が厳しくなりすぎることが知られている．

Bonferroni補正がどのようなケースで厳しくなりすぎるのかを具体的に考えるため，ある培養細胞における遺伝子発現をDNAマイクロアレイで測定する実験を考えてみよう．

[例3] 10,000個の遺伝子の発現を測定できるDNAマイクロアレイを用いて，異なる2つの条件でゲノムワイドな遺伝子発現量の測定を行なう．対照実験として，細胞サンプルになんの処理を施さない実験（以降，C実験とよぶ）を行ない，処置実験として，ちょうど半数（つまり，5000個）の遺伝子の発現を上昇させることがあらかじめわかっている実験（以降，T実験とよぶ）を行なう．C実験とT実験ともに，3枚のDNAマイクロアレイで遺伝子発現量を測定する．また，発現が上昇する5000個の遺伝子を *T*+遺伝子，発現が変化しない残り5000個の遺伝子を *T*0遺伝子とよぶことにする（図4）．

もちろん，このような都合のよい実験は現実にはありえないが，ここでは，有意水準の調整について考えるため単純化した状況を想定している．C実験とT実験とのあいだで各遺伝子の発現量の平均値に差があるかどうかを *t* 検定で検定することにすると，*T*0遺伝子の発現量分布はC実験とT実験で変わらないので，仮説“2つのデータ群が平均の等しい2つの正規分布から得られたものである”が成り立つ．一方，*T*+遺伝子の発現量分布では，T実験での遺伝子発現量の分布が大きいほうにずれていることになる（図5）．

各遺伝子について有意水準を0.01として検定した場合には，発現量変化のない *T*0遺伝子の1%（つまり，50個）が偽陽性となる．もちろん，毎回の実験で正確に50個の偽陽性がでるわけではなく，ここで例として用いている実験をくり返したときに，平均50個の偽陽性が得られるということである．Bonferroni補正では各検定の有意水準として補正前の有意水準を検定回数で割った値（= 0.01/10000）を用いるので，偽陽性のでる確率も1/10000となり，200回の実験をくり返して1個の偽陽性のでる程度となる．しかし，困ったことに，この例の場合には有意水準が *T*+遺伝子の分布よりはるかに高いところになってしまうため，陽性判定になる遺伝子もほとんどなくなってしまう（図5）．もし，*T*0遺伝子と *T*+遺伝子の発現量の差がBonferroni補正で得られる有意水準をクリアするほど大きく離れていれば問題はないのだが，現実に得られるデータはここで例示しているように，分布の一部あるいはほとんどが重なるほど近接しているので，Bonferroni補正では有意水準が厳しくなりすぎて陽性判定が得られないことになる．

では，開き直って，大規模データ解析においては解析者にとって好ましい数の陽性判定が得られるよう適当に有意水準を決めればいいのだろうか？　仮にそうしたとき，得られた陽性判定が偽陽性ではないことをどれだけ強く主張できるのだろうか？　この問題に答えるため，陽性判定に

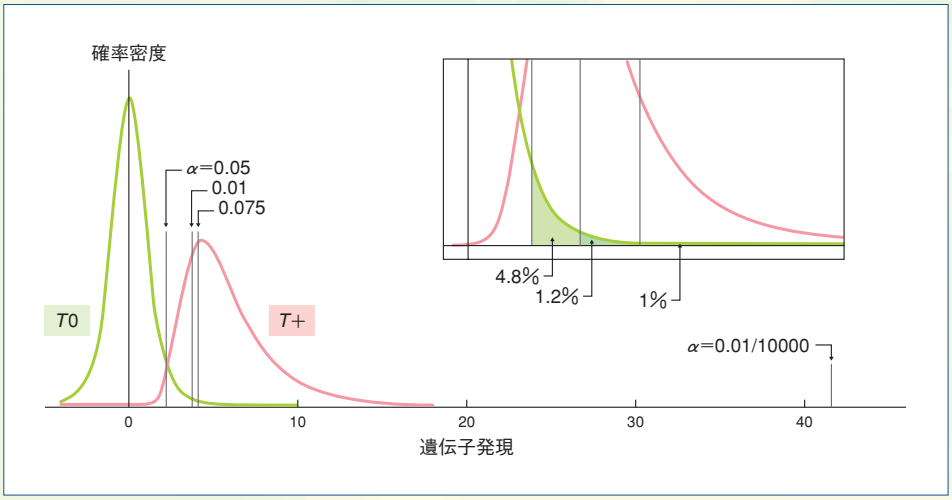


図5　T実験とC実験とのあいだでの *T*0遺伝子と *T*+遺伝子の発現量の差の分布

緑色の線が *T*0遺伝子の発現量の差の分布，ピンク色の線が *T*+遺伝子の発現量の差の分布．縦線は有意水準（α）の対応する発現量の差を示す．α=0.01/10000はBonferroni補正を用いた場合の有意水準．また，α = 0.05では陽性判定データの約4.8%が擬陽性であり，α = 0.01では約1.2%が擬陽性，α = 0.0075で約1%が擬陽性となる．挿入図に，対応する面積を示す．

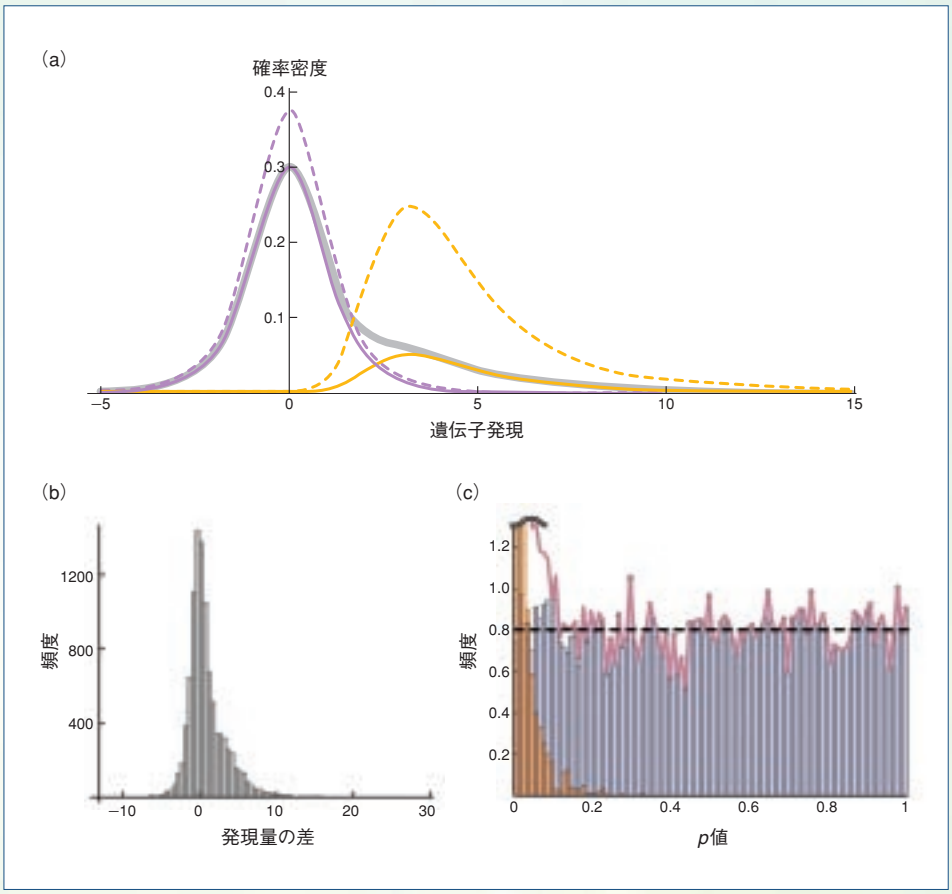


図6　陽性データ分布と陰性データ分布の割合の推定

(a) 紫色の線が陰性データ分布，黄色の線が陽性データ分布で，実線はそれぞれ面積 π_0 と $1-\pi_0$ の曲線，破線は面積1の曲線．灰色の線は陰性データ分布と陽性データ分布の和で，面積は1になる．ここでは，$\pi_0=0.8$ として作図した．
(b) (a) の灰色の線で示した分布にしたがう10,000点のデータ分布．陰性データと陽性データがそれぞれ π_0 と $1-\pi_0$ の割合で混ざっている．実験によって得られるデータ分布を再現している．実際には，このようなデータから π_0 を推定する．
(c) (b) のデータから求めた p 値のヒストグラム．陽性データの p 値と陰性データの p 値の頻度を，それぞれ，橙色と紫色の縦棒で描いた．赤紫色の折れ線は両者の頻度の和を表わす．折れ線の下の面積（全データ数）が1になるようスケールをあわせている．p 値が1に近い領域には陽性データはほとんどなく，陰性データのみになっているため，この領域の y 軸の値が陰性データ分布の割合（π_0）と等しくなっていることが確認できる．

含まれる偽陽性の割合を考えるのが以下で紹介するFDRである．

FDR

Bonferroni補正では，陰性データの分布（さきの例では，$T0$ 遺伝子の発現量の差の分布）のみに着目しながら有意水準を調整することによって，偽陽性数を減らそうとしていた．そして，その結果として有意水準が厳しくなりすぎてしまった．そこで，陽性データの分布（さきの例では，$T+$ 遺伝子の発現量の差の分布）も考慮に入れて有意水準を設定することを考える．具体的には，陽性判定データのなかに混ざっていると見積もられる偽陽性の割合が許容範囲になるよう有意水準を調整する．さきの例では，有意水準を0.05としたときには陽性判定データのうち約4.8％が偽陽性であり，有意水準を0.01としたときにはうち約1.2％が偽陽性，有意水準0.0075でうち約1％が偽陽性になる[*11]（図5）．このように，ある有意水準を決めたとき，陽性判定となる遺伝子のなかに含まれる偽陽性の割合を表わす値をfalse discovery rate（FDR）とよぶ[*12]．もし，ある有意水準を決めたときのFDRが0.01なら，そのときの陽性判定となるデータのなかに混ざっている偽陽性の割合は1％と見積もられるということを意味する．そして，それが解析者にとって許容できるものであれば，対応する p 値を有意水準として採用する．任意の有意水準についてFDRを計算することができるので，各遺伝子の p 値を有意水準として各遺伝子のFDRを得ることもできる．

ところで，さきの例では，陽性データ分布と陰性データ分布が具体的にわかっている状況を想定していた．しかし，現実の実験データの分布には陽性データ分布と陰性データ分布が未知の割合で混ざり合っているので，解析者はFDR

＊11　この例では，2つの正規分布のパラメーターを以下のように定めて用いた．（$T0$ 遺伝子分布，$T+$ 遺伝子分布）として，平均（0，4），標準偏差（1，1），サンプル数（3，3）．これらから，C実験とT実験とのあいだの発現量の差の分布は，$T0$ 遺伝子では自由度4の t 分布となり，$T+$ 遺伝子では自由度4，非心度 $2\sqrt{6}$ の非心 t 分布となる．
＊12　有意水準 t に対応するFDRの式は，FDR (t) = |擬陽性の数| / |陽性判定データ数| = |t 以下の p 値の陰性データ数| / [|t 以下の p 値の陰性データ数| + |t 以下の p 値の陽性データ数|] と書くことができる．

の計算のためその割合を推定する必要がある（図6a, b）．この推定のために利用できる手がかりは，陰性データ分布が帰無仮説にもとづくデータ分布であるという事実である．たとえば，さきの例ではt検定を考えており，“2つのデータ群が平均の等しい2つの正規分布から得られたものである”が帰無仮説であった．そして，帰無仮説にもとづく陰性データのp値は一様に分布するのに対して，陽性データのp値は0に近いところにかたよって分布するという性質を手がかりとすることで，p値のヒストグラムから全データにしめる陰性データの割合を推定できる．具体的には，この陰性データの割合をπ_0と表わすことにすると（このとき，全データにしめる陽性データの割合は$1-\pi_0$で表わされる），π_0はp値のヒストグラムにおいて，0から十分に離れたp値の頻度から推定される（図6c）．

全遺伝子の数にこのπ_0を乗じ，さらに，有意水準であるp値（ここでは，tとする）を乗じることにより，この有意水準tに対応する偽陽性の数が推定される．また，陽性判定される遺伝子の数はt以下のp値をもつ遺伝子の数を数えれば求まるので，FDRの式は，FDR$(t)=\pi_0\times$｛全遺伝子の数｝$\times t/$｛t以下のp値をもつ遺伝子の数｝と書くことができる．

ここで述べたFDRの定義は，Storeyら[1]にもとづくものである．また，彼らは，p値のリストからπ_0を自動的に求めるアルゴリズムを開発し，さらに，p値の減少に応じてFDRも単調減少するよう再定義したq値という値を提案している．彼らによって，p値のリストからπ_0を自動的に求めるソフトウェアがインターネット上に公開されており，これを利用すれば，全遺伝子のp値のリストからq値のリストを容易に得ることができる．

ところで，陽性データが陰性データに比べて十分に少ないと考えられる場合には，よりシンプルに$\pi_0=1$とするのも妥当であろう．この場合，π_0を積極的に推定するよりも，FDRの値は大きく（保守的に）なる．本稿においては，以降，FDRとq値を区別せずにFDRとよぶことにするが，一般的には，q値はStoreyらの定義した指標をさし，FDRは$\pi_0=1$とした保守的な指標として用いられることが多い．

別の方法の実験での確認を

ここまで，多数の統計検定をいちどに実施する多重検定の際には，統計学で伝統的に用いられてきた0.01や0.05という有意水準を用いることができないとする注意にはじまり，多重検定における古典的な対処方法であるBonferroni補正もDNAマイクロアレイ実験に代表されるような大規模データ解析においては実用にならないことにふれ，近年，提案され大規模データ解析において広く用いられているFDRを紹介した．FDRを用いて有意水準を検討した場合には，陽性判定となる遺伝子のなかに含まれる偽陽性遺伝子の割合が見積もられる．このため，解析者は，陽性判定となる遺伝子の数と，そのなかに含まれるであろう偽陽性遺伝子の割合という2つの情報をあわせて考慮して，つぎのステップに持ち込む遺伝子の数を決定できる．さらに，有意水準をどれだけ厳しくしてもFDRが満足に小さくならない場合には，陰性データと陽性データとを分離できていない，つまり，実験もしくはデータ解析に問題があることが示唆されるということも特筆される．

FDRによって大規模データの解析にも有意水準を設定できるようになったが，これまでみてきたような統計的な手法が数学的な取り扱いを容易にするためのさまざまな前提のうえに立脚していることには，あらためて注意を喚起しておきたい．実際，現実の実験データには統計学的な前提を破るものが少なくない．これに対処するための統計的な手法開発からの努力に関してはのちにふれるが，ここでは，なにより統計解析の結果を確認するための検証実験が不可欠であることを強調しておく．とくに，統計解析の対象となった実験手法とは別の手法を用いて検証実験を実施することが望まれる．たとえば，DNAマイクロアレイを用いた遺伝子の発現量変化の研究であれば，陽性判定となった遺伝子について定量的PCR法を用いて発現量変化を確認することで，結果の信頼性は大きく向上する．

もちろん，DNAマイクロアレイ以外でも同様の注意が有効である．たとえば，ゲノム上に散在する哺乳類体内時計の転写制御システムを担う転写因子結合配列を数千カ所予測し，FDRによってそれぞれを評価して，そのなかでもっとも予測スコアの高いいくつかの転写因子結合配列に関して，レポーター遺伝子を作製して培養細胞内でその機能を実験的に検証した報告がある[2]．

おわりに：必ず陽性対照実験を！

本稿では，DNAマイクロアレイによる遺伝子発現量の差の検出を例に，大規模データ解析の問題点と対処方法を解説した．ここまではとくに強調しなかったが，実は，t検定を多数回くり返す際には以下の2つの前提をおいていた．①遺伝子の発現量分布は正規分布にしたがっている．②ひと

つの実験条件につき数枚のDNAマイクロアレイを用いることで各遺伝子の発現量の分散を正確に推定できる．しかし，これら2つの前提は実際のDNAマイクロアレイによるデータではまず成り立たないことが経験的に知られている．この問題に対処するため，たとえば，significance analysis of microarray（SAM）という手法や，empirical Bayes法（経験ベイズ法）などといった方法が開発され用いられているが，このほかにもさまざまな統計手法が現在もさかんに開発されており，万能な解析手法はいまのところ存在しない．したがって，解析手法とデータとの相性を確認しながら解析手法を選択するというプロセスが不可避である．そして，解析手法の選択のためには，実験的な陽性対照（ポジコン）データを準備しておくことが必要である．大規模データ解析において，実験とデータ解析の双方がうまく機能していることを確認できる手段を実験設計の段階で確保しておかなければ，データ解析の段階になって途方に暮れることになってしまう．

さまざまな解析手法があるものの，大規模データの解析において，統計検定をくり返して多数個のp値を得て，そこからFDRを算出する，という流れは，多くの場合で共通して用いられる汎用的な枠組みとなっている．そして，どんな手法を採用したとしても，そののちの検証実験が不可欠であることはすでに述べたとおりである．

最後に，各国の研究者によって現在もさかんに開発が続けられているさまざまな統計的な手法の多くが，Rという統計解析ソフトウェア上で利用できるBioconductorというプロジェクトにまとめられ，インターネットで一般に公開されていることを紹介する．プログラミングや統計に不慣れな人にとって決して敷居は低くないと思われるが，本稿ではわずかにふれるにとどまったq値，SAM法や，経験ベイズ法を含む強力なツールがそろっていることから，非常に魅力的な存在である．また，近年では，関連する書籍なども増えてきている[3]．また，統計解析専用のソフトウェアではないが，数式処理ソフトウェアとして知名度の高いWolfram社のMathematicaも，大規模データ解析において力を発揮する．本稿で用いた例や図は，このMathematicaで計算し作図したものをもとにしている．

文　献

1） Storey, J. D., Tibshirani, R.: *Proc. Natl. Acad. Sci. USA*, **100**, 9440-9445（2003）
2） Kumaki, Y. *et al.*: *Proc. Natl. Acad. Sci. USA*, **105**, 14946-14951（2008）
3） RとBioconductorを用いたバイオインフォマティクス: R. ジェントルマン 他編, 荒川和晴 他訳, シュプリンガー・ジャパン（2007）

山田陸裕
略歴： 2009年 東京工業大学大学院総合理工学研究科 知能システム科学専攻博士課程 修了，2003年より理化学研究所発生・再生科学総合研究センターに在籍，現在，同 基礎科学特別研究員．
研究テーマ： 生命が時間情報など環境情報を獲得する機構に，データ解析と微細流路技術によりせまる．