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. 6

分子生物学，生化学，細胞生物学における統計のポイント

医療統計学の専門家を交えた鼎談

山中伸弥・青井貴之・佐藤俊哉

はじめに

近年，分子生物学や生化学，細胞生物学においても，得られた実験データを統計的に処理して検定を行なうことが，とくに論文執筆などにおいて求められるようになっている．統計とは，実際に得られた個々のデータの集合から，なんらかの規則性あるいは傾向を見いだすことなどによって，真実を把握するための手法であるということもできるだろう．平均気温や株価，野球選手の打率など，わたしたちは日々，テレビや新聞でさまざまな統計に接しているし，その値から“今年の夏は暑い”とか“イチローはよく打つ”などという“データの解釈”を日常的に行なっているのだが，いざ，自分の実験データを統計的に処理するとなると，どのように行なえばよいのか，とたんに自信がなくなる．これは，われわれの多くがこれまで統計というものにきちんと取り組んでこなかったことや，われわれの扱う実験データ数は統計学の教科書で扱われているものより少ない場合が多いことなどの事情によると思われる．

得られた実験データを正しく解釈するためにどのような統計処理を行なえばよいのかが本稿のテーマである．これまでの本シリーズの連載とは趣を変え，今回は鼎談形式とし，多くの読者と同じく統計学に自信のない生命科学研究者である山中および青井が，医療統計学の専門家である佐藤に日ごろの疑問を投げかけた．

Shinya Yamanaka[1], Takashi Aoi[1], Toshiya Sato[2]
[1]京都大学物質－細胞統合システム拠点 iPS 細胞研究センター基礎生物学部門，[2]京都大学大学院医学研究科 社会健康医学系専攻 医療統計学分野
E-mail：yamanaka@cira.kyoto-u.ac.jp
takaaoi@kuhp.kyoto-u.ac.jp
shun@pbh.med.kyoto-u.ac.jp
URL：http://www.icems.kyoto-u.ac.jp/cira/j/index.html
http://www.kbs.med.kyoto-u.ac.jp/

サンプル数が少ない場合に統計処理をする意味はあるか？

山中：佐藤教授，本日はお忙しいなかありがとうございます．『蛋白質 核酸 酵素』での“正しい知識が捏造を防ぐ”という連載の第6回目で，統計に関する正しい知識を身につけ，誤った報告をしてしまわないためにはどうすればよいのか，というテーマについて教えてほしく訪問しました．わたしたち，分子生物学，生化学，細胞生物学を専門にしている者は，これまであまりきちんと統計処理をしてこなかったことが多いのですが，近年は，やはり統計は大事できちんとやりなさい，ということが，われわれの分野でもっとも権威のある学術雑誌でも強調されるようになりました．たとえば，*Nature Cell Biology* 誌の投稿規定には，統計に関してかなり具体的なガイドラインが示されています（http://www.nature.com/ncb/pdf/gta.pdf）．

佐藤：はじめまして．実は，ぼくは薬剤の臨床試験や疾患の危険因子の抽出などをおもな研究対象にしていて，動物実験のデータってあまりみたことがなく，そんなにくわしくはないんですよ．ですから，質問に対して正確な答えをもちあわせているわけではないかもしれませんが，

表1　細胞移植治療の有無と運動機能スコア

	個体番号	測定値
細胞移植群	A	5
	B	8
	C	1
	D	7
対照群	E	5
	F	6
	G	5

もってこられたテーマについて，いっしょに考えていきましょう．

山中：よろしくお願いします．さっそくですが，まず，このデータをみてください（表1）．これは仮想のデータですが，ある疾患モデルのサルに幹細胞移植治療を行なった場合の運動機能を点数で表現したものです．サルですからあまりたくさん使えず，細胞移植群が4匹で対照群が3匹となっています．わたしたちがふだん扱う実験データはこのようなものが非常に多いのです．ここでまず疑問なのが，この4とか3とかいうサンプル数で，そもそも統計処理をする対象になるのか？　ということです．実際には，なんらかの統計処理をしていますが，このぐらいのサンプル数のときだと，意味のあることだといえるのでしょうか？　あるいは，なんらかの計算式にあてはめて値をだして満足しているだけなのしょうか？　これが以前からすごく疑問だったのですが，いかがでしょうか？

まずはサンプル数に応じたデータの示し方を

佐藤：いろいろな動物実験データや解析をみせてもらって，対象となっている動物の数が少ないということはよく承知しています．で，そのときにですね，統計学の教科書をみると，基本的にはデータ数が多くあることを前提としていますから，何かを表示するときに，平均をだしたり標準偏差をだしたりします．でも，4匹とか3匹とかの世界で平均をだしたり標準偏差をだしたりとかするのは，そんなに意味がないと思うんですね．そこで，もうちょっと違う表示の仕方，平均とか標準偏差ではなくて，たとえば，4とか3くらいの数だったら，具体的な数字をならべてグラフにするだけでも十分に効果があるかどうかわかる場合があると思います．ですから，とにかく統計的に処理をすればいいんだということではなく，まずは記述の仕方，つまり，データにあったグラフの表示とか表の表示とかということを考えたほうがいいと思います．

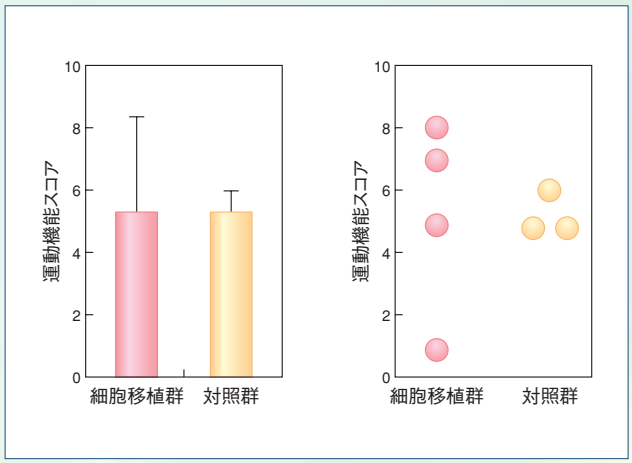


図1　データの示し方

(a) 平均値±標準偏差で示した棒グラフ．
(b) 同じデータから作成したドットプロット．

山中：なるほど．*Nature Cell Biology*誌の統計ガイドラインにも，サンプル数が3以下の場合はすべてのデータをプロットしなさいと書いてあります（図1）．

佐藤：そうでしょうね，統計学者の立場からはそういうコメントをすると思いますね．

山中：サンプル数が5以下になる場合は全データをプロットして，そこにエラーバーをつけてもよいと書いてあります．ただ，明らかにオーバーラップのないような効果だったらいいのですが，たとえばこの4とか3くらいの数であれば，さきほどいわれたように，あまり統計処理をする意味はないということでしょうか？

佐藤：相当に違いがないと統計的に差があるとはいえないと思いますね．誰がみても違う明らかな効果があれば検出できますけど，微妙な違いはわからないでしょうね．

必要なサンプル数は効果の差の大きさに依存する

山中：では，逆に，治療群と対照群とのあいだに明らかに差のある場合は，これくらいのサンプルでも統計処理をすると有意な結果となるのでしょうか？

佐藤：いま，ぼくがよく関係しているのは臨床試験とか疫学研究とかなんですけど，そこでは非常に小さい差を検出しようとしています．いま，治療薬なんかもだいぶん頭打ちになってきていて劇的な効果のあるものがないですから．そうすると，臨床試験に参加する患者さんがたくさん必要ですし，それから，環境リスクについても，

たとえば1.2倍だけ死亡が増えるというようなものですと，たくさんの方に参加してもらう必要があるんです．昔，抗生物質ができたときのことを考えると，抗生物質は使えば必ず治りますから，そういう意味では，抗生物質が効くか効かないかなんていうのは，別に対照群がなくても，統計処理をしなくても，明らかになるものは明らかになるんですね，ただやはり，そんなに明白にわかるものがだんだん少なくなってきたこと，それから，どうしても生物学的なメカニズムによって日内変動や日間変動などがありますから，それをこえて効果があるかどうかをみようと思うと，やはり対照群を設けて統計処理をすることが必要になってくるんですね．

山中：必要なサンプル数は効果の差の大きさに依存するということでしょうか？

佐藤：そのとおりです．ですから，臨床試験でも非常に治療効果の高い場合には，対象者の数が少なくても統計的な有意差はでますし，逆に，臨床的には意味があるけれども治療効果の小さいものを検出しようと思うと，非常にたくさんの患者さんに参加してもらわなくてはならない．だから一概に，こういう研究だから何人とか何匹いなくてはならないとかではなくて，みたい効果の大きさ，それと，その指標が生物学的にどのくらい変動するのかということも考えて決まってくると思うんですね．

山中：なるほど，わかりました．

少ないサンプル数に適した統計の手法は？

山中：つぎの質問に移ります．*Nature Cell Biology*誌の統計ガイドラインに，サンプル数が少ないときは，それに適した統計処理をしなさいと書いてあるんですが，そういう方法はあるんでしょうか？

佐藤：これもむずかしい問題で，統計学の教科書に書いてある方法というのは，検定にしても推定にしても，みな，近似なんですよ．どういう近似かというと，対象数が非常に大きいときには，いろいろなものが正規分布に近似的になる，ということを使って，いろいろな計算をしているんです．しかし，対象数が多くないとその近似は成り立たないので，対象数が少ないときには注意をしなさい，対象者数が少ない，あるいは，対象数が少ないときには，もっと正確な方法を使ってやらなくては駄目ですよ，ということが書かれているんだと思います．でも，さっきもいったように，対象数が少ないと，いろいろな統計処理をしても小さな差はみえてこないので，正確な方法を使ったからよいというものではないんですね．もちろん，だから近似を使ってよいということでもないし，対象数が少ないときは注意をして，あまり過度にまとめた情報をださないで，3匹以下だったら必ず全部をプロットしなさい，という，個体ごとのデータがわかるように開示すればいいと思うんです．少ない対象数の場合には，正確な方法を使えるときには，近似的な方法よりも正確な方法を使ったほうがよいといわれています．それも，やはり基本的には，ある程度の数がそろって，その傾向をみるというのが統計的な考え方です．

山中：具体的にいうと，それはノンパラメトリックな方法[*1]でやるということでしょうか？

佐藤：そうですね．ノンパラメトリックな方法もその一部ですね．

サンプルではなく母集団の分布が重要

山中：関連した質問になりますが，正規分布をしているかどうかという検定方法があると思いますが，その結果にかかわらず，サンプル数が4とか5とかいう少ない数の場合，そもそも正規分布ということ自体が前提として成り立たなくなるように思うのですが？

佐藤：正規分布しているかどうかと，データ数の問題は，あまり関係がありません．たとえば，平均値を*t*検定[*2]で比べる場合には，動物を何匹か選んできて実験することになるんですけど，その背後にある動物の非常に大きな集団（母集団といいます）を考えたとき，何かの特性が正規分布しているかどうかが問題で，別に手元にきた100匹とか5匹の動物（これをサンプルといいます）が正規分布しているかどうかは問題ではないんです．正規分布から選ばれた代表なのかどうかが問題で，もし，背後にある集団で何かの特性が正規分布をしていれば*t*検定というのは正確な方法となります．だから，*t*検定は近似ではないんです．ただ，背景にある集団が正規分布していないと，必ずしも正確な方法にはなりません．そうでない場

＊1　**ノンパラメトリックな方法**：母集団の分布が正規分布するという（パラメトリックという）などの仮定をおかずに，比較的広い範囲の分布のもとでも妥当な統計的方法．

＊2　***t*検定**：“2群の正規分布する母集団の平均値の差が等しい”という帰無仮説を検定する方法であるが，母集団の分布が左右対称の場合，よい性質をもつ．

合でも，t検定は非常によい性質をもっていることがわかっているので，平均値を比較したいという場合には，t検定を使ってもノンパラメトリックな方法に比べてそれほど大きな問題は起こりません．ノンパラメトリックな方法を使えばよいかというと，背後にある分布に仮定をしていないのでよい方法ではありますが，t検定だったら，その代わり，背後にある分布が正規分布をしていれば差があるときに検出できる可能性が高くなります．背後にある分布を仮定しないと，ほんとうに差があるときに差があるといえる威力が落ちてしまうのです．だから，その損を覚悟のうえでノンパラメトリックな方法を使うというのであれば，それでかまわないと思います．

青井：理論的に母集団が正規分布をしているんであれば，たとえば，サンプルが5例あって，それが明らかにかたよった5例でも，それはたった5例だから母集団の一部をかたよってとってきてしまっただけと考えて，パラメトリックな方法を使うということはありうるのですか？

佐藤：別にかまわないと思います．その5匹というのは，背後にある集団からなんらかの意味で代表性をもっていると考えられる．明らかにかたよって選んだのではなく，選んだけどたまたまかたよっていたということ，それはわからないですから，しょうがないです（図2）．

山中：では，たとえば，ある血圧の薬を投与した人と投与していない人を調べると，当然，投与していない人たちの血圧は正規分布するのは予想されますが，投与した人たちの血圧も違う山で正規分布する可能性が高い．つまり，両方が正規分布するんですが，山はシフトしますよね？

佐藤：はい．平均値がシフトします．

山中：そのことはかまわないのですか？

佐藤：基本的に，平均値の検定にしてもノンパラメトリックな方法にしても，何を調べているかというと，全体の値がなんらかのかたちで分布するわけですけれども，それを要約するにはいろいろなやり方があるんです．それで，その代表的なものが平均と標準偏差ということになります．位置を示すのには，平均値のほか，中央値や最頻値があります．また，ばらつきがどのくらいかを示すものに，標準偏差や範囲，四分位範囲があります．そして，平均値の差の検定にしてもほかの検定にしても，分布の位置が同じかどうかということを調べていて，ばらつきについては，位置がかわっても基本的には変わらないという仮定をおいているんですね．ですから，血圧でしたらおそらく，ほんとうは血圧が下がるだけでなくて，ばらつきも大きくなるとか小さくなるとかということがあると考えられますけども，基本的には，血圧の平均値が薬によってどのくらい変化したのか，つまり，位置がずれたか，ということを調べているのが平均値の検定なんですね．

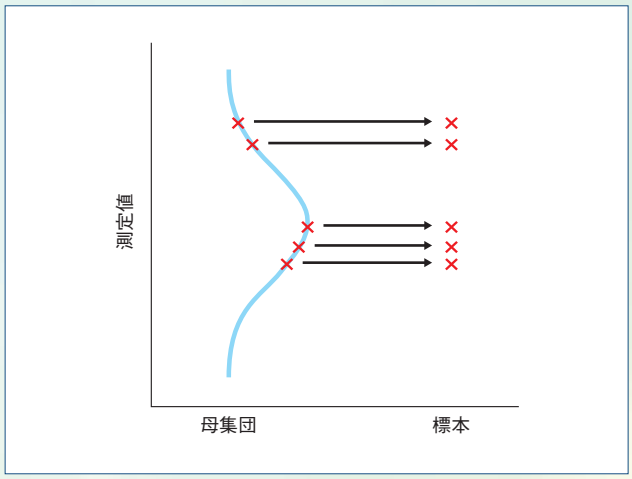


図2　母集団と標本

山中：それは普通のt検定においてもでしょうか？

佐藤：そうですね．正規分布が仮定できればt検定でかまわないと思います．正規分布が仮定できなくても，左右対象の分布がもともとの分布に近いものだったら，ほとんど問題ないですね．それから，もっというと，まったく正規分布でないような分布をしていても，対象数が多くなればt検定でほとんど問題なく検定することができます．だから，今回みたいにデータ数が少ないというときは問題になってくるとも思います．

データの分布に応じた表現

山中：さきほど話がでた，データに応じた位置やばらつきの適切な表現方法はあるでしょうか？　たとえば，サンプルが少ないときは標準偏差ではなくて範囲を使いなさい，というような？

佐藤：はい．ただしこの問題は，別にサンプルが少ないからということではありません．たとえば，さきほどふれた平均と標準偏差を何で使うかというと，もしデータが正規分布にしたがっていれば，正規分布というのは，平均と標準偏差が決まれば全部が決まってしまうんですよ．2つの値を決めれば全部決まってしまう，非常に性質のよ

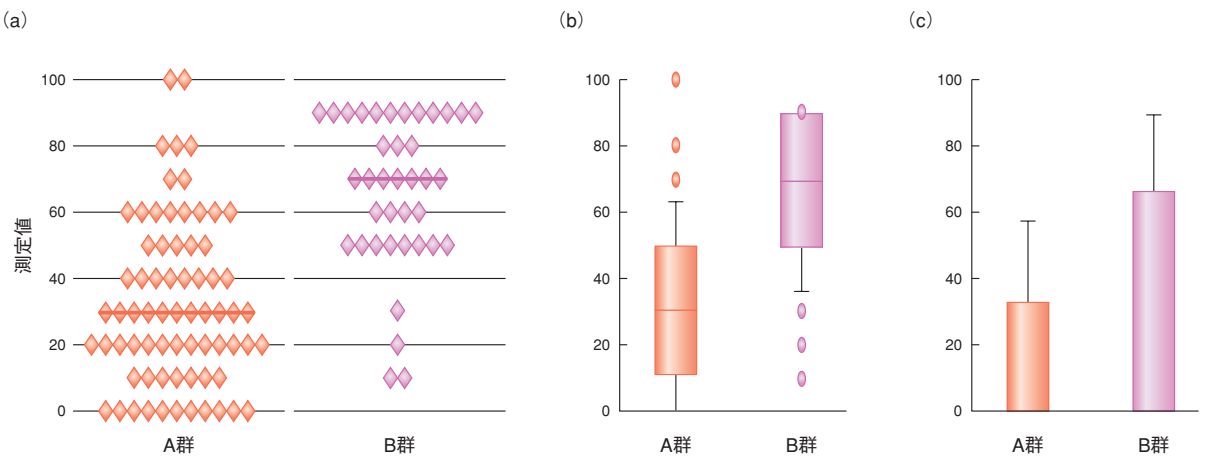
い分布ですね．ところが，ほかの分布をしていると平均と標準偏差だけではなくて，たとえば，右にひずんでいてその程度がどのくらいかとか，正規分布より尖っていてその具合がどのくらいかとか，分布を決めるのにたくさんのパラメーターが必要になってきます．たとえば，生存時間分布などは左右対称にならないんですね．長生きする人がいますから，右に伸びたようなカーブで．そのとき，平均と標準偏差だけでいいかといったら，たぶんよくないです．ですから，ぼくも生存時間などを調べるときは必ず中央値と範囲，つまり，最小値と最大値を示すことにしています．データ数が多いからといって，平均と標準偏差だけでよいということではないんですね．データが左右対称に近い分布をしていれば，平均と標準偏差だけでほとんど問題ないと思います．左右対称の分布ではなかったら，むしろ平均とか標準偏差を使うよりも，中央値とか範囲とか，そういうものを使ったほうがデータの要約としては適切だと考えられます（図3）．

青井：データが正規分布にしたがわない場合に，サンプル数が4とか5とかだったらプロットするのがいちばんよいとして，もうちょっと多い，何十くらいのデータだとボックスプロットが図示するにはよいと思っています．それに対応する本文の記載ですが，中央値（最小値-最大値）という記載をよくみかけます．でも，範囲よりむしろ四分位範囲を示したほうが，本文を読んでいるだけでむしろ分布をイメージしやすいようにも思うのですが？

佐藤：もちろん，それでもかまわないですよ．

青井：でも，そのような記述をあまりみかけないのが不思議な気がします．

佐藤：いや，ぼくにとっては，中央値と四分位範囲をみたからといって，どんなばらつきになっているかは逆にイメージしにくいから，最大値と最小値のほうがいいと思っています．とくに，データ数が少ないと，はずれたデータがあるだろうから，そのほうがいいと思います．図については，ボックスプロットで書けば，はずれたデータ



(d)

	A群	B群
平均値	32.1	66.3
中央値	30	70
最頻値	20	90

(e)

	A群	B群
標準偏差	25.2	22.9
範囲	0–100	10–90
25パーセンタイル	10	50
75パーセンタイル	50	90

図3　同じデータから作成した3種類のグラフと種々の代表値

(a) ドットプロット．
(b) ボックスプロット（箱ひげ図）．
(c) 棒グラフ（平均値±標準偏差）．
(d) 位置を示す代表値．
(e) ばらつきを示す代表値．

もわかるし，それから四分位範囲もわかるので，もっと使われてもいいと思いますよ．

山中：標準偏差が意味のある値にならない分布の場合には，標準偏差のエラーバーをつけるのは正しくないと考えてよいでしょうか？

佐藤：はい．基本的に，標準偏差は正規分布ないしは左右対称に近い場合には有効ですね．サンプル数が小さい場合にも，母集団の分布がどうか，ということが大切です．たとえば，サンプル数が5前後ぐらいというのは統計処理の行なわれることも多いのですが，母集団が正規分布ないし左右対称の分布をしていると考えられる場合，平均や標準偏差を計算する意味はないことはないと思うんです．

多群の比較：分散分析と多重比較

山中：多くのグループを比較するときにはANOVA[*3]を使うべきで，t検定をくり返したらだめです，とよくいわれます．でも，たとえばA，B，C，Dという薬があってAだけすごい効果があるときに，じゃあ，対照群とA処理群だけを取り出してt検定をして報告したくなるのですが，これは許されることでしょうか？

佐藤：それは駄目ですね．医薬品の開発のときにはどうしているかというと，計画の段階で事前に決めるんです．これとこれの差は確実にみる，あとは参考にしますと．それでしたら1個だけだしても問題ないですし，あるいは，全部みるんだったらもう最初に全部みるということをあらかじめ設定しておいて，そのうえで多重性を調整する方向で全部を比較するということを宣言します．

検証的な研究と探索的な研究

佐藤：われわれはよく，検証的な研究と探索的な研究ということをいいます．医薬品などでとくに大規模な臨床試験をするときには，検証的なステージですから確実な結果をだす項目はなにかを事前にきちんと決めます．あとは大規模な試験になりますから，ひとつだけ調べるのももったいですから，つぎにつなげるためにこういう項目も調べます，あるいは，主要な結果をサポートする項目がこれなので，これについても調べます，ということをならべるんですけれども，それはもう，おまけなんですね．

表2　細胞移植治療の有無と運動機能スコアの時系列変化

		移植前	10日目	20日目	30日目
細胞移植群	A	4	5	7	9
	B	5	5	8	10
	C	4	5	7	8
	D	5	5	6	8
対照群	E	4	5	5	7
	F	4	4	5	6
	G	5	5	6	7

おまけなので，多重性の調整もしない．ただ，みて差がついていても，強い結論をださずに，あくまでもつぎの研究につなげるための情報とする．その代わり，主要な項目に定めた1個については，確実な結果がでるように統計的に対象者数の設定をします．どのくらいの差があったら確実に有意になり，どのくらい差がなかったら有意にならない，という対象者数を計算するんです．そういうやり方をしますから，あまり多重性の問題は起こりません．しかし一方で，探索的に調べるときにも，やっぱりいろいろ調べて，どうだったかということをいいたいと思います．いろいろなパターンが考えられますが，たとえば，時系列のデータを例に考えます（表2）．こういった時系列のとき，10日目で差があったかどうか，20日目で差があったかどうか，30日目で差があったどうか，ということを調べることがよくあります．そうすると，10日目，20日目，30日目と平均値の検定をしていって，20日目が有意だったら，20日目に効きましたという結論にはならないんです．知りたいのは，最終的に効果があるかどうかということですから．

大切なのは“その実験で何を明らかにしたいのか？”

佐藤：つまり，その実験の目的，何をいいたいのかということで，ほんとうに10日ごとに検定しなくてはならないのか，ということが問題になります．もし，その薬がたとえば，ある程度の期間の服薬ののちしばらくたってから効果がでるものだとすれば，30日目最後のところで1回，差があるかどうか調べるということだけですし，あるいは，トレンドとして明らかに差がでてくるということであれば，個々のポイントではなくて，変化の仕方自体がもう違うんだよ，ということがわかるような検定方法

＊3　ANOVA（analysis of variance，分散分析）：2群の平均値の差のt検定を多群の平均値の差に拡張したもの．

を使えばいいわけです．この時系列で比較することの大きな目的が10日目，20日目，30日目でそれぞれ平均値の差を検定することではないので．だから，多重比較というよりも，何回も検定しても意味がないということですね．それから，グループ比較のほうは，別に2グループずつの比較をt検定で行なってもいいのです．それはただ，有意水準5％で，たとえば3回も4回も比較をすると，有意水準1回1回が5％だとしても，全体としてみるとそれが過剰になってしまうので，そこを調整しましょう．それが多重比較の考え方なんですね．いちばん単純にいったら，Bonferroni法[*4]で何分の1かに調整して検定をすれば，別に2群ずつ総あたりで検定してもかまわないんです．しかし，そうするとちょっと損をする．つまり，ほんとうは差があるのにそれを検出できない可能性があるんです．ほんとうに2群ずつの比較が必要なんだったら別に多重比較の方法があって，その方法を使ったほうがほんとうに差があるときの検出力が上がります．目的がほんとうに2グループずつ総あたりで比較してどれがまさっているのかみつけたいというなら，t検定のくり返しでまったくかまわないと思います．ただ，有意水準のことについてちょっと注意しましょうね，というだけの話です．そういう意味からいうと，探索的な研究だったら別に有意水準を調整する必要性もないんです，．その結果，差があったからといって検証するわけではないんですから．また，注意をしてつぎの検定に進みましょう，と．一方，検証的にものをいうためにはかなり有意水準を厳しくしておかないと，ほんとうは差がないのにまちがって差があるとするαエラー（第1種の過誤）が大きくなってしまう．有意水準が大きくなってしまうので，そこは注意しましょう，ということです．ただ，有意水準を小さくすると，今度は差があるときに差を検出する力が小さくなり，ほんとうは差があるのにまちがって差がないとするβエラー（第2種の過誤）の可能性が大きくなりますから，バランスをどこでとるかということが重要です．

山中：たとえば，薬を12種類，試したいけれども，比べたいのは互いの状態ではなくて，対照と比べて効果があるかどうかをみたい．そのときは全部で11の組合せになりますが，それぞれを対照群とt検定するということはあるのでしょうか？

佐藤：ありますし，そういう場合に適切な多重比較法もあります．総あたりの組合せで比較をするんではなくて，ひとつ対照群があって，それといくつかの試験群とを比較するというDunnett法という多重比較の方法も提案されています．それを使ったほうが，対照群とそれぞれを3つとか4つ検定するよりも検出力が上がります．ただ問題は，ほんとうに12種類の薬剤を対照群と比較して検定することが科学的に意味のあることなのか，ということになると思います．

事前の研究計画と結果の正確な記述

山中：表2のような時系列の例で，何日目に効果がでるのかまったくわからないという前提で，10日ごとに3ポイントで評価したところ，実際に効果がありそうなのは最後のポイントだったとします．その場合，30日目のデータだけを取り出して，移植前のデータと統計比較することは許されないということでしょうか？

佐藤：そうですね．探索的な研究であっても，結果がでるまえに解析方法も決めておかないといけません．いろいろな解析を試みることはいいんですけど，いろいろやったらすべてを報告しないと，よかったところだけだしているのではないかといわれますから．

山中：“捏造を防ぐ”という点では，もう実験のデザインの段階で統計方法は決めておいて，そのとおりに解析を行なった結果はこうでした，というべきなのですね．

佐藤：そのとおりです．とくに，検証的な実験の場合はそうしたほうがよいと思います．

山中：ただ，報告の仕方として，“実験全体としては有意差がでなかったけれども，30日目で上がる傾向にあったので，こことここだけを比べたら，それは有意差がでた．そこは有望なので，今後，さらなる追試が必要である”というような書き方にするのはどうでしょうか？

佐藤：それはまったく問題ないです．まさにいま，臨床試験はそうなっていて，事前に計画した解析と，あとからやった解析で事前に計画していなかった解析は分けて報告しなさいといわれています．事前に計画で設定した解析は信頼性が高いけれども，あとから結果をみて解析した結果は信頼性が低いからそれは分けて，論文にまとめるときも分けて書きなさい，と指示がきますね．やってはいけないというわけではなくて，せっかくやった実験と

＊4　**Bonferroni法**：検定の多重性を調整する方法のひとつで，n個でセットとなる仮説を検定する場合，個々の仮説検定の有意水準を5％/nとするもの．

か研究ですから，いろんな情報を有効に使いたいというのは当然のことだと思うんです．むしろ，何もしないほうがいけないことだと思うんですね．ただし，いろいろみた結果と，事前にやろうと思ったことはしっかり分けて報告する．厳しい人だと，論文も分けろという人もいるぐらいです．ただ，同じ実験，同じ研究のなかで2つに分けて論文をだすというのはちょっとまずいですから，ひとつの論文のなかで，これはあとからデータをみて思いついてやった解析である，ということを書いて考察すれば別に問題はないと思います．

山中：すごくわかりました．自分は大学院のときには薬理学を研究していましたが，そのときは，動物を10匹使ってこういうデザインでやって，こういう統計処理をやる，というのを必ず決めてからそのとおりに解析したのですが，ポスドクのときに分子生物学の研究室にいったらあまりそういうことをやっていなくて，実験結果をみてからもう1回やろうか，ということが多いように思いました．

佐藤：それは，たぶんいろいろな背景があって，臨床試験だと，うまくいかなかったからもう1回やろうって，できないじゃないですか．だから，やっぱり慎重に計画をたてて，ということがわりと根づいたと思うんですけども，細胞レベルとか分子レベルになってくると，もちろん実験はたいへんだと思いますけれども，駄目だったからもう1回やろうよ，というのがわりとできてしまうので．

山中：事前にしっかり計画をたてることは，正しいデータの解釈だけでなく，研究費の節約にも役立ちますね．

はずれ値をどう扱うか？

山中：関連する問題ですが，実験データのなかでほかとはかけ離れた値を示すもの，つまり，はずれ値について，いつも頭を痛めます．ある測定値が人的なサンプル採取のミスと考えた場合，それを除外したら有意差がでたとしても，除外したあとの解析だけを示すのは，さきほどの教えてもらったように不適切に思います．一方で，除外した値は明らかにおかしくて，それを除けば知りたいことの真の答えがわかると考えられるときは，両方を示したらよいような気がするのですが．

佐藤：はずれ値の問題はわれわれの領域にもたくさんでてきているんですけど，基本的には，誤りであることのはっきりした証拠がないかぎり省くべきではない．あるいは，身長が100メートルとかいうありえない値は削除するしかないですけれども，測定のまちがいが記録に戻ってわかったとか，測定者に聞いてわかったとか，そういうことがないかぎり，それは削除してはいけないと学生にも教えています．しょうがないですよね，そういうデータが入ってしまったら．これを除外したいという気持ちはわかるんですけれども，あまりよくないですね．せいぜいできることは，いわれたように，そのデータを含めた結果と，除外した結果との両方を示すということですが，それも，両方の結果が同じだったら実験結果を支持することになりますが，違ったらやはり，実験自体の信頼性が疑われることになります．

山中：10匹ずつ測定しようと思って，たまたま2匹だけが変なことになってしまったとき，それを除外したら差がみえるとします．しかし，さきほどいわれたような明らかな理由がない場合は，慎重に実験をやりなおすしかないのですね．

佐藤：それしかないと思いますね．なぜそういうデータがでてしまったか，原因を追究して，そういうことが起こらないような条件設定をすることが大切です．

ノンパラメトリックな方法の多重比較検定

山中：ノンパラメトリックな方法で分散分析を行なったあとの多重比較については，統計学の教科書をみてもあまり明確な記載がみられない気がするのですが．

佐藤：多重比較の考え方というのは，あくまで考え方なので，別にパラメトリックな方法であろうがノンパラメトリックな方法であろうが，適応できるわけです．いちばん単純にいうと，Bonferroni法には，t検定であろうが中央値の検定であろうが，全部，比較する数で0.05を割った値を使えばよいのです．それと同じように，総あたりで平均値を比較するというTukeyの方法があるんですけど，この方法はノンパラメトリックな方法でも計算することができます．

山中：Scheffeの方法というのはいかがですか？

佐藤：この方法は古くからありますが，ノンパラメトリックな方法に対する考え方もできるはずです．いま，かなりの手法がいろいろな統計ソフトウェアに入っていますから，できるんじゃないかと思いますね．

青井：そもそも，ノンパラメトリックな方法の多重比較というのは統計学の専門家のなかでもまだ議論があるということを耳にしたことがあるのですが．

佐藤：とくに複雑な多重比較だとパラメトリックな方法の場合でもそうです．2群比較だと勝った負けたがはっきりしているから，検出力を高くしたいというのは比較的簡単なんです．でも，3群以上になると，どれとどれが勝って，どれとどれが負けたか，じゃあ，ほんとうの真実の差を検出するというのは何なのか，ということがすごくわかりにくくなってしまうんですね．

統計的有意≠生物学的に意味があること

山中：ほかに何か，統計学の専門家からみて，みながしている誤解などあれば教えてください．

佐藤：よくあるのは検定の偏重です．有意差がでたらよかった，というのがあって，とくに生物医学分野での共通の誤解といったら変ですけれども，検定の結果って，多くの対象数を集めたら絶対に有意になるんですね．身長が1ミリ違うというのも，1000人ずつくらい対象者を集めてきたら絶対に有意差がでますから．統計的に非常に小さいp値になったということと，臨床的もしくは生物学的にそれが意味のあることだというのは別なんです．もともと統計的有意を英語で“significant”というんですが，検定が発明されたのは19世紀の終わりごろなんですよ．そのころの英語の意味といまの英語の意味が違っているらしくて，昔，significantとは，なにかいつもと違うことが起こっているから注意しましょう，いつもと違うからちょっと注意してみたほうがよいですよと，いうだけだったんです．20世紀に入ってから，significantというのが，何か重要なことが起こっている，というような意味に変わったんですね．で，それから，検定の結果significantだったということが，非常に意味のあることがわかったという誤解になって，それがさらに，検定の考え方が医学領域に入ってきたときに，統計的に意味があることは医学的にもイコール意味があるんだ，というように誤解されて広まっている，という感じだと思います．だから，統計的に有意だったといっても，ふつうとは違うことが起こっているよ，ということであって，それがほんとうに生物学的に意味のあることなのか，あるいは，なにか系統的なかたよりでいつもと違うことが起こっているのか，検定では区別できないんですよ．そのことがいちばん重要なのであって，検定の結果でp値が5%より小さいかどうかということよりも，こういった効果を調べるのであったらどのくらい差があるのかとか，信頼区間で調べたらどのくらいの，この実験の規模ではどのくらいの精度なのかという，効果の大きさと効果の推定精度をきちんと報告するのが重要なのだということが，もっと認識されるべきだと思います．

データを正しく表現するために

山中：短時間で統計のすべてを教えてもらうのは無理ですけれど，今回の鼎談で再認識したのは，あらかじめしっかりデザインして，解析方法まで決めてから実験を行ない，そして，結果はすべて報告すべきであるということです．$N = 10$と決めたら報告も$N = 10$であるべきであって，それはちゃんと*Nature Cell Biology*誌の統計ガイドラインにも書いてあります．サンプル数が減った場合は，なぜ減ったのか，きちんと書くことが基本ですね．

佐藤：臨床試験でもそれが問題になっていて，臨床試験の結果をどのように論文に書くかというガイドラインまでできているんですよ．最初に何人必要だったかという数から，実際に何人集まったのか，最終的には何人を比較したのか，どうして途中で数が減ったのか，全部わかるように図を描いて示すことが求められています．最近まで，なかなかそういう認識にならなかったんですね．

山中：そうですね．ここで適当というか，都合よくデータを除外していくと，偽造や捏造になってくるかもしれないわけですね．今日はほんとうに勉強になりました．ありがとうございました．

おわりに

青井：佐藤教授の研究室を訪問するあたって，質問のリストを持参した．Yes/No，あるいは，“この場合はこの検定法を使う”といった一問一答式の回答，いいかえれば，小手先の解決法を授けてもらうことをイメージしていたのである．しかし，教えてもらった話の多くは，重要性をこれまで十分に意識してこなかった，われわれが統計を行なうにあたっての土台となる根本的な考え方であった．しかも，その内容は“研究の目的が何かを明確にして，それにそった適切な方法を選択する”ことや“実験をはじめるまえに研究計画をたて，結果を正直に記載する”“得られた複数の生データをもっともよく表わすような図の作成や記述を行なう”といった，日常，われわれが行なうあらゆる実験においてつねに心がけなければならない，いわば，自然科学の方法のPrincipleともいうべきことば

かりだった．統計処理においてもこれらが重要であるということを今回の鼎談ではじめて意識したという事実こそが，われわれの分野における統計に関する問題の大本ではなかったかと感じた．

山中：佐藤教授が最後にいわれた，“significant”という英語は“いつもと違う，注意すべき”という意味であったのが“重要な”という意味に変わった，ということは目から鱗であった．実験結果が統計的に有意な変化を示したとき，それは“注意すべき”結果であって，学問的に“重要である”または“真実である”とはかぎらないことを肝に銘じるべきである．数年前，ある遺伝子のノックアウトマウスを作製したとき，最初に生まれてきた20匹くらいのホモ変異マウスはすべて雄であった．性決定にかかわる遺伝子か！と色めきたったが，100匹ぐらい解析すると雄雌ほぼ半々となった．最初の数十匹だけを解析すればまちがいなく統計的に有意であろう．その段階でその遺伝子が性決定に役割をもつと報告してしまえば，捏造や偽造ではないが，まちがった報告となってしまう．このような落とし穴に落ちることを防ぐためには，異なるアプローチでの検証，たとえば，cDNAの導入でレスキューされるか，siRNAでも同様の表現型が認められるか，などが必要である．正しい統計解析をすることの重要性と同時に，統計的な有意差を盲信しないことの重要性も，佐藤教授から教えてもらった．

佐藤：今回の鼎談では，自分自身，実験データについてはあまり経験がないので，なにを聞かれるのかドキドキヒヤヒヤものでした（笑）．話をしているうちに，背景は異なりますが，自分が専門としている疫学研究や臨床研究の領域でも問題となっている医療統計学の誤用や誤解とよく似ていることがわかってきて，少し安心しました．“データの分布に応じた表現”のところのボックスプロットの話などは，データをどのようにまとめたらいいのか，ほんとうに平均と標準偏差だけでいいのか，という疑問に対し，Tukey（多重比較のTukeyです）らが“探索的データ解析”の方法のひとつとして提案したものです．また“significant”の意味については，Salsburgが著書“The Lady Tasting Tea”（Henry Holt，2001，邦訳：統計学を拓いた異才たち，日本経済新聞社，2006）のなかで述べています．“データを正しく表現するために”のところのガイドラインは“CONSORT声明”（http://www.consort-statement.org/home/）というもので，臨床試験の結果をどのように論文に書くべきかが定められており，ほとんどのメジャーな医学雑誌ではこのガイドラインをみたすことを投稿規程に盛り込んでいます．とくに，対象者数の変遷についてはフローチャートにすることが強く推奨されていますので参考になると思います．最後に，今回は自分が話を聞きましたが，本来は，実験データを専門とする統計学の専門家がいるべきです．製薬企業の研究所などにはそういう人が何名かいるようですが，大学や研究機関にも統計家のポストを設けて実験家といっしょに研究ができれば，実り多いのではないかと思います．

鼎談のテープ起こしをしてくれた山中研究室秘書室に感謝します．

山中伸弥

略歴：1987年 神戸大学医学部 卒業，1989年まで国立大阪病院臨床研修医，1993年 大阪市立大学大学院医学研究科 修了，米国Gladstone Institute ポスドク，1995年 同Stafff Research Investigator，1996年 大阪市立大学医学部 助手，1999年 奈良先端科学技術大学院大学遺伝子教育研究センター 助教授，2003年 同 教授，2004年 京都大学再生医科学研究所 教授を経て，2007年 京都大学物質ー細胞統合システム拠点 iPS細胞研究センター センター長．

佐藤俊哉

略歴：1986年 東京大学医学部保健学科 助手，1991年 統計数理研究所 助教授を経て，2000年より京都大学大学院医学研究科 教授．