Japanese Psychological Review
2016, Vol. 59, No. 1, 123–128

統計学の現場は一枚岩ではない

三　中　信　宏[1,2]

[1] 農業・食品産業技術総合研究機構農業環境変動研究センター
[2] 東京大学大学院農学生命科学研究科

The actualities of statistics are not monolithic

Nobuhiro MINAKA[1,2]

[1]Institute for Agro-Environmental Sciences, NARO
[2]Graduate School of Agricultural and Life Sciences, The University of Tokyo

The recent controversy over statistical data analyses sheds a light on a number of cases of abuse of statistical procedures. In this essay some practical aspects of statistical analyses, mainly in agricultural research, are discussed. During the past century eminent researchers, including K. Pearson, R. A. Fisher, J. Neyman, and E. S. Pearson, have established the theoretical basis of modern mathematical statistics, e.g., experimental design, sampling distributions, and hypothesis testing. Some users in psychology, agronomy, etc. might be liable to commit misconduct in statistical analysis. Of course while they are responsible for what they have done, they must understand not only the proper use of statistical methodology but also the characteristic of each science.

Key words: statistical inference, experimental design, hypothesis testing, *p*-value, abduction, QRPs
キーワード：統計的推論，実験計画法，仮説検定，*p* 値，アブダクション，QRPs

1. はじめに：統計分析の現場を振り返る

本特集が編まれる発端のひとつにもなった，「*p* 値」をめぐるアメリカ統計学会（The American Statistical Association：ASA）の声明（Wasserstein & Lazar, 2016）は，統計分析のさまざまな現場でいまも広く用いられているある判定基準の“誤用”を枚挙した。統計学界で大きな影響力をもつこの学会があえて発表したこの声明の警鐘は，研究分野の壁を越えて，またたく間に科学者コミュニティーに反響していった（たとえばBaker, 2016）。

長年にわたって主として農業実験分野での統計分析の現場に接する機会が多かった私個人の経験を振り返ると，確かにその声明に指摘されているような統計的データ解析の手法や基準の“誤用”はそれを意図するかしないかに関係なくさまざまな場面で数多く見られた。実験者が納得できる統計的結果が出るまで手段を選ばない不適切な行為は，最近では「*p* 値ハッキング（*p*-hacking）」と呼ばれるようになったが，生態学では以前から「ゆーい差決戦主義」（久保，2003, 2012）と呼ばれていた。私も統計学の講義や研修では「*p* 値バンザイ突撃戦」なる表現を使うこともあった。

本特集では社会心理学や実験心理学での統計分析の現状に絡めた問題点の指摘がなされている。しかし，それらは統計分析の根幹に関わる問題提起であり，他の研究分野でも同様な指摘ができるだろう。それと同時に，ユーザーが統計分析を用いて何をやろうとしているのかは必ずしもひとつではないことも見えてくる。たとえば，本特集の中心的テーマである「再現可能性（replicability）」についていえば，確かに実験系の科学では得られた結果が再現できるかどうかは重要なことかもしれない。しかし，非実験系の科学では結果の再現性よりもむしろきちんと推定できているか，まっとうに説明できているかどうかの方により重きが置かれるだろう。

以下では，私が見聞したさまざまな統計学的

“誤用”を踏まえて，その背後にひそむものに目を向けたい。

2. 統計学の使用と誤用：農業試験研究の場合

私は仕事柄，農業試験研究機関の研究員を相手に，実験計画法（experimental design）についての講義や演習，場合によっては個別コンサルティングを行ってきた。この実験計画法の基本的な理念と技法は，創始者である数理統計学者Fisher（1926）がイギリスのロザムステッド農業試験場に在籍していたときに開発したものである（参照：Box, 1978；芝村，2004；Giuditta, 2015）。農業実験の現場ではいまでも実験区の配置をする際にFisherの実験計画法の原理は必須である（三輪，2015；三中，2015）。

正規分布母集団からのサンプリングを前提とするFisher流の実験計画法は，もっと洗練された線形統計モデルが幅を利かせている現代にあっては，いささか時代遅れの“レガシー”な統計分析とみなされてもしかたがないかもしれない。現場の統計ユーザーにとっては，広がり始めている“新しい統計学”について知る機会がないだけのことが多いので，いったんその味を体験すれば次の一歩を踏み出す動機づけとなるだろう。その後押しをするのはもちろん私の仕事のひとつである。

その一方で，過去に実施された研究の系譜を伝承しなければならない現場のニーズを考えるなら，“レガシー”な統計手法であっても適切な使用法と誤用の回避を知ることはとても重要である。母集団からの少数の標本であっても，適切な実験計画を組めば，正確な統計的推論を行うことが可能であることを示したFisherの理念は現代でもなお通用する。その点からいえば，不必要に大量のサンプルを抽出して決着をつけようとする態度は，Fisherに先行するK. Pearsonの時代（Porter, 2004；芝村，2004）への“先祖返り”を髣髴とさせる。

推測統計学の基盤を築いたFisherは，帰無仮説を明示することである有意水準のもとでの検定を実行するという方針を据えた。その方針をさらに一歩進めて，対立仮説と対比することで意思決定としての統計的検定の枠組みを完成したのがNeymanとE. S. Pearson（Neyman & Pearson, 1933）だった。現在のユーザーが基礎知識としてもっている（はずの）“レガシー”な統計分析はさかのぼれば一世紀近く前にすでに確立されていたということだ。

FisherそしてNeymanとE. S. Pearsonが目指した実験計画法と推測統計学の中心理論の根幹は，実験や観察を始める「前」に，実験区の割り付けを完了し，帰無仮説と対立仮説を設定し，仮説検定のための有意水準を決めることにある。Fisher（1926）が提示した実験計画法の三原則は次の通りである：

1）「反復実施」：同一実験処理を複数回実施することにより，その処理にともなうばらつきを評価する。
2）「無作為化」：実験処理区のランダムな配置をすることにより，背景要因によるデータへの体系的な影響を偶然誤差化する。
3）「局所管理」：実験場所を適切にブロック分割することにより，ブロック内の実験環境の均一化をはかる。

いったん実験や観察が開始されたならば，それらの初期設定を変えてはならないし，事後の統計解析は事前の実験計画に忠実に沿わなければならない。Fisherはイギリス王立統計学会の会長就任講演で，「実験終了後に統計学者に相談を持ちかけるのは，統計学者に，単に死後診察を行って下さいと頼むようなものである。統計学者はおそらく何が原因で実験が失敗したかという実験の死因について意見を述べてくれるだけであろう」（Fisher, 1953；Rao, 1997: 183から引用）と述べたほどである。

ところが，私が見てきた農業試験研究の現場では必ずしもそうではない。たとえば，本来ならば「反復（replicate）」は別々の実験区から複数回抽出しなければならないにもかかわらず，同一の実験区から複数個のサンプルを抽出したもので代用するという「擬似反復（pseudoreplicate）」の誤用がきわめて多く見られると指摘されている（Hurlbert, 1984；山村，1999）。擬似反復を使えばたくさんの実験区を用意する必要がないからだ。これはもちろん「反復実施」の原則に反する。「無作為化」に違反して，無作為化すべき実験区をちゃんと無作為化しなかったという初歩的なミ

スもいまなおある。また，乱塊法のブロックの切り方があやふやな事例も少なからず見受けられる。まちがったブロック設置は「局所管理」の原則に抵触する危険がある。

このような実験計画法上の“誤用”を生む原因には，実験者がもともと実験区配置の理論を知らなかったとか，（農業試験場ではよくあることだが）前任者が実施した試験設計をそのまま継承せざるを得ないという情状酌量の余地がある場合もある。しかし，その一方で，得られたデータから何とか“有意”な検定結果を導き出すために故意に行われる“不正”の手口もいろいろ見聞きした。上述の「擬似反復」のほかにも，つごうの悪いデータに「外れ値（outlier）」という主観的なレッテルを貼って解析から除外するという事例もある。さらには，多要因実験で高次の交互作用項を恣意的に誤差とみなすことで，自由度を荒稼ぎして，検定結果を有意にもちこむというような“裏ワザ”が農業試験ではときどきある（「p 値ハッキング」の一例）。あるいは，実験前に仮定した統計モデルとは異なる分散分析を事後的に適用してしまうという事例もある。得られたデータを前にしてモデルそのものを操作するというこの“誤用”は「HARKing」そのもので，その動機は検定結果を有意にしたいという思惑である。このようなダークゾーンの「QRPs」は農業試験研究では相当前からあったものと推測される。

今回の特集論文では，心理学研究におけるさまざまな統計学の“誤用”とそれらへの対策が論じられているが，農業試験研究を“反面教師”としてさまざまな“統計学的不正”の手口を知っておくことは何かの役に立つかもしれない。少なくとも，研究分野を問わず，実験観察の目的が「5%レベル有意性の星」あるいは「p 値の小数点以下の0の個数」のみにあるとき，さまざまな“不正”の手口が編み出されるのはやむを得ないことだろう。だからこそ，そのような現状を憂えたアメリカ統計学会はあえて声明まで出したにちがいない。

3. 統計的推論の目標は何か？：強い推論と弱い推論

確かに，統計データ解析の個々の手法を解説したり，統計計算ソフトウェアの使い方を伝授することは有用ではある。しかし，私の経験から言えば，統計学的な「ものの考え方」の理解を促す方がもっと重要でありしかもはるかに難度が高い。

どんな統計手法にも必ずそれが生み出されるに至った具体的な問題状況があったはずであり，さらにその背後には理念的・哲学的なバックグラウンドがあったはずである。しかし，現代の統計学者の多くはそのような統計学史や統計学哲学にはほとんど関心がないように見える。もちろん，一般の統計ユーザーのほとんどにとっては，手持ちのデータを適当な統計ツールをつかって計算できさえすれば満足であり，やっかいなめんどうくさいことに深入りする気はさらさらないにちがいない。

私的な憶測としていえば，統計分析ソフトウェアのインターフェイスが快適になればなるほど，ユーザーはものを考えなくなるようだ。たとえば，アメリカ統計学会の声明（Wasserstein & Lazar, 2016）に挙げられている「p 値の誤用」リストを見ると，p 値はある仮説の「真実性」「証拠」「効果量」などのいずれにも関係がないと書かれている。しかし，これらは古典的な統計学をちゃんと勉強していれば犯すはずのないまちがいではないだろうか。

本特集のいくつかの論文で挙げられている，将来に向けての学会あるいは学会誌レベルでの「対策案」は，要するに統計ユーザーが“誤用”や“不正”をしないようにという意図で提案されているのだろう。グッド・ラック！

しかし，統計ユーザーの個人レベルに目を向けたとき，もう少しやっかいな問題が浮上する。私が方々の大学や農業試験場で統計の講義を行ったとき，よく訊かれるのは「どんな統計手法を使えば“正しい答え”が出せますか」という質問である。おそらく，その質問者にとっての統計手法は“真実”を見通す“水晶球”のようなものなのだろう。データを統計分析にかけて“真実”が転がり出れば“当たり”というわけだ。そのとき，「統計を使ってもほんとうのことはわかりませんよ」と身も蓋もない答えを返すと，相手は多くの場合かなり落胆してしまうようだ。

もちろん，統計データ解析は“真実”を見つける術などではない。確かにK. Pearsonは19世紀

末以降に大流行した論理実証主義の空気をまともに吸い込んでいただろう（Porter, 2004）。大量のサンプルを取れば“真実”がつかめると夢を描いたとしても不思議ではない。しかし，その後の現代統計学が展開した20世紀は，科学哲学も同時に発展した時代でもあった。既知のデータから未知への統計学的推論をいかに進め，その結論をどのように解釈するかは，ただ統計数学だけの問題ではなかったはずである。

Raoは次のように述べている：「特定のものから一般化を行うという規則によって作り出された知識は，不確実なものであるが，ひとたびその中に含まれる不確実性を数量化すれば，それは，種類は異なるが，確かな知識となる」（Rao, 1997；芝村，2004: 123から引用）。統計学が得意とする“不確定性”すなわち偶然のばらつきの定量化は推論の上で強力な武器となる。では，統計学が目指している推論とはいかなる性質を帯びているのだろうか。ここで，統計的推論のもつ認識論的な考察が必要になる。

統計的仮説検定を取り上げよう。古典的な仮説検定の方法論も時代によって変遷があった。たとえば，Fisherは対立仮説を設定せずに帰無仮説を検定しようとしたが，Neyman–Pearsonは帰無仮説に対置する対立仮説を仮定したというちがいがある（Hacking, 1965；Barnett, 1999）。Neyman–Pearsonの仮説検定の枠組みによれば，あるデータのもとで仮説検定を行ったとき，検定統計量が棄却域に入れば，帰無仮説を棄却するという意思決定を行う。これは古典統計学を学べば誰もが叩きこまれる基本事項のひとつだ。しかし，この仮説検定の枠組みはそれが確立された1930年代と変わらないままずっと継承されている。

Royallは，このNeyman–Pearsonの枠組みそのものに問題があると指摘した：「統計学という分野はそれが取り組むべきある重要問題の解決を怠ってきた。その問題とは，得られた観測値は，どのようなときに一方の仮説を支持するが，他方の仮説は支持しないといえるのかという問題である。すなわち，その観測値が対立する仮説のうちの一方を支持する証拠とみなしてもいいのかということだ」（Royall, 1997: xi）。この問題が議論されてこなかった理由について，彼はこう言う：「過去半世紀にわたって統計理論は意思決定（decision-making）パラダイムに支配されてきた。1930年代のNeymanとPearsonの研究以来，統計学の根本問題は対立する行為のいずれを選択するかの意思決定問題として定式化され，データを証拠（evidence）とみなしてはこなかった」（Royall, 1997: xi）。

いまから80年前に定式化されたNeyman–Pearsonの意思決定パラダイムに対抗する，データを仮説に対する“証拠”とみなす新たなパラダイム（Royallは「尤度パラダイム」と称する）の提唱は，統計的推論の科学哲学に大きく踏み込むことになった。なぜなら，意思決定パラダイムが帰無仮説と対立仮説の命運を分ける絶対的な基準を置くのに対し，尤度パラダイムは仮説間の証拠（すなわち尤度）による相対的な重みづけをするだけで，仮説の受容や棄却の意思決定を伴わないからである。

ここではデータと仮説との認識論的関係性が問われている。Sober（1988）は，提示された仮説の“真偽”を得られたデータによって判断しようとする立場を「強確証／強反証」と名づけた。一方，データを証拠として仮説の相対的な“支持”の強弱を判定する立場は「弱確証／弱反証」と呼ばれる。

Neyman–Pearsonパラダイムから離れて統計的推論を考察するとき，「アブダクション（abduction）」という推論の形式は注目に値する（Josephson & Josephson, 1994；Lipton, 2004；Walton, 2005）。アブダクションという推論は，データを説明するために立てられた仮説の“真偽”を問わない。むしろ，同一のデータを説明しようと競合する複数の対立仮説の間で，データを証拠とする相対的な“支持”の順位を踏まえ，その時点でもっともよい仮説を選び出す。

このアブダクションの推論様式は次のように定式化できる（Josephson & Josephson, 1994；三中 2006, 2009）：

前提1）観察データDがある。
前提2）ある仮説Hはデータ Dを説明できる。
前提3）H以外のすべての対立仮説H’はHほどうまくDを説明できない。
結論）仮説Hを最良として受け入れる。

アブダクションの手順を上のように整理すると，対立する他の仮説とのデータ＝証拠に基づく

相対的比較が決定的であることがわかる。仮説の“真偽”を問わないアブダクションには推論の終わりがない。将来的に新しく追加されたデータあるいは新たに立てられた仮説との比較により，現時点での推測が覆される可能性はつねに残されている。このように，アブダクションとは果てしない推測の連鎖である。

統計的推論をアブダクションのためのツールであると考えるならば，個別の科学の性格に応じてうまく使い回すことができるのではないだろうか。たとえば，Sober は生物の進化的系統発生を推定する系統学（phylogenetics）というある研究分野での仮説（系統樹）の相対的判定のために上述の「弱確証／弱反証」という用語を提唱したが，これらは他の分野にも適用できる一般性をもっているだろう。

4. おわりに：ふたつの科学のはざまで

科学は一枚岩ではない。一方には，仮説の真偽が実験によって白黒をつけることができる実験系の科学もある。他方には，系統学のように，直接的な観察や実験がまったくできない歴史叙述科学（historiographic sciences: Tucker, 2004）のような科学もある。実験科学ならば結果の再現可能性が問われることは十分にありえるだろうし，それに対して綿密な実験計画のもとに結論を得ることはきっと可能だろう。しかし，非実験科学ではそもそも再現可能性という概念そのものを適用することが原理的に無理なので，歴史的事象の痕跡からアブダクションによって過去を復元することをいつまでも続けていかねばならない宿命にある。佐倉統のコメントにもあるように，科学研究における再現可能性を過度に重視することの弊害についても一考すべきだろう。

もちろん，実験科学と非実験科学とは峻別できるわけではけっしてない。Laudan は歴史科学と非歴史科学を対置して次のように言う：「信頼の置ける知識を得るための方法に関しては，歴史科学と非歴史科学という分け方にたいした意味はない。確かに，過去のものやできごとは直接的には観察できない。しかし，非歴史科学が対象としているものやできごとであっても直接観察できない場合は少なくない。そういう障害を克服しようと努力しなければならないのはどんな科学でも同じである」（Laudan, 1992: 65）。限られたデータから統計的推論を行うとき，われわれは自分の手がけている科学がはたしてどんな性格をもった科学であるのかをつねに問い続ける必要があるだろう。

さて，心理学ははたしてどのようなタイプの科学を目指していくのだろうか？

引用文献

Baker, M. (2016). Statisticians issue warning over misuse of *P* values: Policy statement aims to halt missteps in the quest for certainty. *Nature*, *531*, 151.

Barnett, V. (1999). *Comparative statistical inference, 3rd edition*. Chichester: John Wiley & Sons.

Box, J. F. (1978). *R. A. Fisher: The life of a scientist*. New York: John Wiley & Sons.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, *33*, 503–513.

Fisher, R. A. (1953). The expansion of statistics. *Journal of the Royal Statistical Society, Series A (General)*, *116*, 1–10.

Giuditta, P. (2015). The emergence of modern statistics in agricultural science: Analysis of variance, experimental design and the reshaping of research at Rothamsted Experimental Station, 1919–1933. *Journal of the History of Biology*, *48*, 301–335.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge: Cambridge University Press.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*, 187–211.

Josephson, J. R., & Josephson, S. G. (Eds.) (1994). *Abductive inference: Computation, philosophy, technology*. Cambridge: Cambridge University Press.

久保拓弥（2003）樹木・森林生態学「よく出る」誤用統計学の基本わざ　生物科学，*54*, 188–192.

久保拓弥（2012）データ解析のための統計モデリング入門：一般化線形モデル・階層ベイズモデル・MCMC　岩波書店.

Laudan, R. (1992). What's so special about the past? In M. H. Nitecki & D. V. Nitecki (Eds.). *History and evolution* (pp. 55–67). Albany: State University of New York Press.

Lipton, P. (2004). *Inference to the best explanation, 2nd edition*. London: Routledge.

三中信宏（2006）系統樹思考の世界：すべてはツリーとともに　講談社.

三中信宏（2009）分類思考の世界：なぜヒトは万物を「種」に分けるのか　講談社.
三中信宏（2015）みなか先生といっしょに 統計学の王国を歩いてみよう：情報の海と推論の山を越える翼をアナタに！　羊土社.
三輪哲久（2015）実験計画法と分散分析　朝倉書店.
Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289–337.
Porter, T. M. (2004). *Karl Pearson: The scientific life in a statistical Age*. Princeton: Princeton University Press.
Rao, C. R. (1997). *Statistics and truth: Putting chance to work, 2nd edition*. River Edge: World Scientific. 藤越康祝・柳井晴夫・田栗正章（訳）（2010）統計学とは何か：偶然を生かす　筑摩書房.
Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Boca Raton: Chapman & Hall/CRC.
芝村 良（2004）R. A. フィッシャーの統計理論：推測統計学の形成とその社会的背景　九州大学出版会.
Sober, E. (1988). *Reconstructing the past: Parsimony, evolution, and inference*. Massachusetts: The MIT Press. 三中信宏（訳）（2010）過去を復元する：最節約原理，進化論，推論　勁草書房.
Tucker, A. (2004). *Our knowledge of the past: A philosophy of historiography*. Cambridge: Cambridge University Press.
Walton, D. (2005). *Abductive reasoning*. Tuscaloosa: The University of Alabama Press.
Wasserstein, R. L., & Lazar, N.A. (2016). The ASA's statement on *p*-values: context, process, and purpose. *The American Statistician*. doi: 10.1080/00031305.2016.1154108
山村光司（1999）土壌肥料学における数理統計手法の応用上の問題点：4. Pseudoreplication と繰り返し測定　日本土壌肥料學雑誌，*70*, 84–89.

― 2016. 4. 24　受稿，2016. 4. 25　受理 ―