はじめに

ダレル・ハフはその有名な著書『統計でウソをつく法』*1 の最後の章で,「医療の専門家に関わるもの」や科学研究室・大学によって裏づけられたものには信じる価値があると述べている。無条件に信頼を置くのではなく，メディアや政治家よりもきっと信頼が置けるだろうということだ（何しろ，ハフの本は，政治やメディアで用いられるまぎらわしい統計的なまやかしで埋め尽くされているのだ)。だが，訓練された科学者による統計に文句をつける人はほとんどいない。科学者は，政敵に対して用いるような攻撃手段ではなくて，知性を追い求めるものだ。

統計的データ分析は科学の基礎だ。気に入った医学誌の中からランダムに 1 ページを開けば，*t* 検定，*p* 値，比例ハザードモデル，傾向スコア，ロジスティック回帰，最小二乗当てはめ，信頼区間といった統計に圧倒されるだろう。統計学者は，最も複雑なデータセットの中から秩序と意味を見いだすという巨大な力を持つ道具立てを科学者に提供し，科学者は，大喜びでこうした道具立てを受け入れてきた。

しかし，科学者は，統計教育を受け入れてこなかった。そして，科学に関する大学学部課程の多くで，統計の訓練はまったく求められていない。

---

*1　訳注：ダレル・ハフ（Darrell Huff, 1913-2001）は米国の著述家・ジャーナリスト。統計の専門家というわけではなかったが，その著書『統計でウソをつく法』(*How to Lie with Statistics*）は英語版だけで 150 万冊以上売れた[9]。

1980年代以降，研究者は，評判の高い査読*2付きの科学文献に，多数の統計に関する誤謬と誤解があることを示してきた。また，多くの科学論文――半分以上かもしれない――がこうした誤りにはまっていることを見いだしてきた。多くの研究が，検定力の不足によって，探求しようとしていることが発見できなくなっている。多重比較とp値の解釈の誤りによって，多数の偽陽性が引き起こされている。融通無碍（ゆうづうむげ）なデータ分析によって，何も存在しないところに相関関係を発見することが簡単になってしまっている。そして，不適切なモデルを選ぶことによって，重要な結果が歪（ゆが）んでいる*3。ほとんどの誤りは，特別な統計の訓練を受けていないことが多い査読者や編集者によって見逃されている。投稿を吟味する統計学者を雇う学術誌はほとんどないし，正確に評価するために必要な統計の詳細を十分に書いている論文はほとんどないからだ。

問題は不正が行われていることではない。問題は貧弱な統計教育だ。これは，研究上の発見で公刊されたもののほとんどが誤っているかもしれないと一部の科学者が結論づけるのに十分なほど，貧弱なのだ[1]。一流の学術誌には，論評記事や編集者からの論説が定期的に出ていて，統計に関する基準をより高いものにし，さらに精査するように求めている。だが，こうした懇願に応じている科学者はほとんどおらず，学術誌が定めた標準はしばしば無視される。そして，統計に関するアドバイスは，統計の教科書――これは誤解を招くことがしばしばある――だけでなく，さまざまな学術誌における論評記事や科学者には理解しにくい論文にまき散らされている。このため，ほとんどの科学者は，統計の実践を簡単に改善できないのだ。

現代の研究の方法論が複雑であることは，統計の幅広い訓練を受けていない科学者が，自らの専門分野で公表された研究のほとんどを理解できない可能性があるという事態をもたらす。例として，医学分野を見てみよう。標準的な統

*2 訳注：科学者が学術誌に論文を載せようとして，学術誌の担当者に論文を送りつけたとしても，すぐにその論文が掲載されるわけではない。学術誌側は，掲載を決める前に，論文が学問的な意味で問題がないかを調べる。このことを査読（peer review）という。査読で問題がないと判断されてはじめて論文として学術誌に掲載され，公刊される。

*3 訳注：検定力の不足の問題は第2章に，多重比較とp値の解釈の問題は第4章に，融通無碍なデータ分析の問題は第9章に，不適切なモデルを選ぶことの問題は第8章に詳しい説明がある。

計の入門講義を１つしか受けていない医師の知識は，『ニュー・イングランド・ジャーナル・オブ・メディシン』（*New England Journal of Medicine*）に掲載された研究論文のうち，およそ５分の１しか完全に理解できない程度のものだ[2]。ほとんどの医者はそれよりも受けている統計の訓練が少ない。多くの医学研修生*4は，必修科目として統計を学ぶのではなく，輪読会や短期講習で非公式に統計を学ぶ[3]。医学生に教えられている内容がしっかりと理解されないことはしばしばある。医学分野でよく使われている統計手法に関するテストに対する医学研修生の正答率は平均して 50% 以下だった[4]。研究に関する訓練を受けている医学校の教授陣ですら，正答率は 75% に満たなかった。

状況は非常によろしくない。統計知識に関する調査を作成した人ですら，調査質問を練りあげるのに不可欠な統計知識を欠いているぐらいなのだ。このため，つい先ほど引用した数字も誤解を招くものになっている。というのも，医学研修生に対して実施された調査には，p 値の定義を問うという多肢選択式問題で，４つの誤った定義しか選択肢にないという問題が含まれていたのだ[5]。ただ，多少は大目に見ることができるかもしれない。多くの統計の入門書も同様に，この基本的な概念の定義があやしかったり間違っていたりするからだ。

科学研究の計画を立てる人が十分に注意して統計を用いなければ，何年もの作業と何千ドルもの資金を費やして，答えようとした問題に答えられない可能性すらある。心理学者のポール・ミールは，以下のように不満を述べている。

> 一方では，科学の論理の考慮ということにひるまず，現代の統計的仮説検定の「正確さ」に喜々として頼るようなやる気に満ちあふれた研究者が，公刊された文献の長いリストを作り出し，教授に昇進していく。こうした人物は，長く残る心理学知識の根幹に対してほとんど貢献していない。こうした人物の本当の位置づけは，強力だが不毛な知的放蕩（ほうとう）者に過ぎない。その浮かれた人生の道には，陵辱（りょうじょく）されたおとめたちが長い列を作っているが，科学

*4 訳注：米国では，医学校（medical school）で医学に関する基礎知識を学んだ後に，病院で医学研修生（medical resident）として実践的な研修を受けて，正式の医師になるのが通常のコースとなっている。日本の医学教育で言えば，研修医が比較的位置づけが近いものになるだろう。

に関して生き残った子どもはいないのだ[6]。

知性が欠如しているとほとんどの科学者を非難するのは不公平かもしれない。ほとんどの学術分野は，p 値の誤った解釈以上のものにのっとって成立しているからだ。だが，こうした誤りは，現実世界に非常に大きな影響を及ぼしている。医学における臨床試験は，私たちの健康管理を左右するし，強力な新しい処方薬の安全性を決定する。犯罪学者は犯罪を減らすためのさまざまな方法を評価するし，疫学研究者は新しい疾病（しっぺい）の速度をゆるめようとする。マーケティング担当者やビジネスマネージャーは商品を売る最善の方法を見つけようとする。こうしたことは，つまるところ，統計に行き着く。ダメな統計学，なのだ。

何が良いか悪いかについて医者が決定しないということに不満を述べたことがある人は，誰でもこの問題の範囲を理解するだろう。私たちは，今や，何らかの食品や食事や運動が有害かもしれないと主張するニュース記事を拒否するような態度になっている。数か月後の必然的な第二の研究を待つだけの話だ。その研究は正反対の結果を示しているだろう。ある著名な疫学研究者は「私たちは急速に社会のやっかいものになっている。人々は，もはや私たちのことを真剣に取り合ってくれない。そして，人々が真剣に受け取ってくれるときは，私たちは，意図せず，有益なことより有害なことを多くなしているのかもしれない」と述べている[7]。私たちの勘（かん）は正しい。多くの分野で，最初に出てきた結果は後から出てきた結果と矛盾する。刺激的な結果を早く頻繁に公刊しようとする圧力の方が，追加の証拠で支持され，慎重に確かめられた結果を公刊しようとする責任よりも強いようなのだ。

ただ，そんなに急いで判断しないようにしよう。いくつかの統計に関する誤りは，単に資金や資源が不足したことによって起こっている。1970 年代半ばに，ガソリンと時間を節約するために，米国で運転手に赤信号での右折を許すようにした運動について考えてみよう。このようにしても変更前に比べて衝突事故が増えることはないという証拠は，統計的におかしいものだった。そして，この後すぐに見るように*5，赤信号での右折を許すようになった結果，多くの人命が失われた。交通安全の研究者を妨げた唯一の要因は，データの不足だった。もし，より多くのデータを集め，より多くの研究を実施するための資

金があれば，そしてさまざまな州の独立した研究者の結果を比較対照する時間があれば，真実は明らかになっていただろう。

「無能で十分説明できることを悪意のせいにするな」というハンロンの剃刀（かみそり）の話がある一方で，「ウソ，くそったれなウソ，そして統計」式の結果で公刊されたものがある*6。製薬産業は，薬に効果がないことを示す研究を公刊しないことで証拠を歪ませる誘惑に特に駆（か）られているように見える*7,8。後から文献を評価する人は，薬に効果がないことを示す公刊されなかった 8 個の研究を知らないままに，薬に効果があることを示す他の公刊された 12 個の研究を見つけて満足するだろう。もちろん，薬に効果がないとする研究はたとえ投稿されたとしても，査読付きの学術誌で公刊されることはないかもしれない。つまらない結果に対する強い偏見があるために，効果がなかったと述べる研究は決して公にならず，他の研究者がそれを見ることができない状況がもたらされるのだ。データが欠けていることと公刊の偏りは，重要な問題に関する認識を歪め，科学への災いとなっている。

正しく行われた統計ですら信じることができない。統計の手法や分析で使用可能なものが多すぎるため，研究者がかなり自由にデータを分析できるようになっている。そして，「データが吐くまで拷問（ごうもん）する」ことはとても簡単だ。統計ソフトが提供しているさまざまな分析をどれかが興味深い結果を出すまで試し，そうした結果を出した分析こそが最初からやろうとしていた分析だったと

---

*5 訳注：第 2.2.2 節「赤信号での誤った方向転換」を参照のこと。

*6 訳注：「ウソ，くそったれなウソ，そして統計」という言葉は，統計で人をだますことを示した警句に由来している。この警句は，「ウソには 3 種類のものがある。ウソ，くそったれなウソ，そして統計だ」（There are three kinds of lies: lies, damned lies, and statistics）という形で用いられる。本文ではこの警句について触れることで，人をだますために統計を悪用した研究があることを示している。なお，この警句は一般には 19 世紀後半の英国の首相ベンジャミン・ディズレーリによるものだと知られているが，本当はディズレーリのものではないらしい。

*7 原注：製薬産業での統計に関する災いに興味がある読者は，ベン・ゴールドエーカーが書いた『悪の製薬』（*Bad Pharma*［Faber & Faber, 2012］）を楽しめるかもしれない。この本を読んで，私の血圧は統計的に有意な増加を見せた。

*8 訳注：『悪の製薬』は 2015 年に青土社より忠平美幸・増子久美による和訳が出版されている。

いつわるのだ。超能力なしに，ある公刊された結果が，データへの拷問で得られたものかどうかを判断することはほとんど不可能だ。

理論があまり数量に基づくものでなく，実験の計画が難しく，手法があまり標準化されていないような柔らかい分野では，このような自由が付け加わることが顕著な偏りを引き起こす[8]。米国にいる研究者は，キャリアを進めるために，興味深い結果を生み出して公刊しなくてはならない。大学などの研究職で空いているものはわずかで，このわずかな数の職を求める競争は激しいものになっている。こうした競争があるために，科学者は統計的に有意でない結果を生み出すだけのデータを数か月あるいは数年かけて収集したり分析したりすることはできないのだ。こうした科学者は，悪意がなくても，データから許されるところより自分の仮説に都合が良い方向に誇張された結果を生み出す傾向がある。

これからのページで，こうしたありふれた誤りとその他もろもろのことを紹介していきたい。誤りの多くは，公刊された大量の文献の中にはびこっている。何千もの論文について，その報告に疑いの目が向けられているのだ。

近年，多くの人が統計の改革を呼びかけている。そして，当然，そうした人の中でも，問題対処の方法として何が一番良いかということについて，意見の相違がある。p 値については，まぎらわしく混乱を招くことがしばしばあると説明する予定だが，これの使用を完全にやめるべきだと主張している人がいる。他には，信頼区間に基づく「新しい統計学」を提唱している人もいる*9。さらには，より解釈がしやすい結果を出す新しいベイズ的手法に切り替えるべきだと提案している人もいる。また，今教えられている統計で大丈夫だが，用いられ方が良くないと信じている人もいる。こうした立場はどれも見るべきところがある。私としてはどれか１つの立場を選び取って，この本で推奨するつもりはない。むしろ，私は，現役の科学者によって現在用いられている統計に着目している。

*9 訳注：「新しい統計学」(new statistics) を主張している人物として，ジェフ・カミングがいる。カミングは，心理学研究を念頭に，信頼区間を重視した統計分析をすべきだとしている[10]。