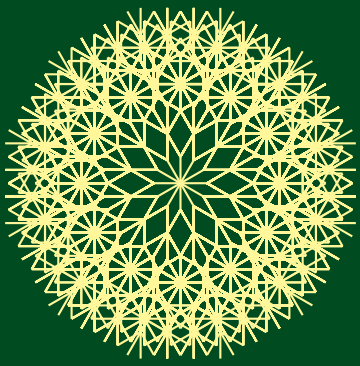
# ダメな統計学

## Statistics Done Wrong



アレックス・ラインハート〔著〕
西原 史暁〔訳〕

# ダメな統計学

2014 年 12 月 28 日　日本語版公開

アレックス・ラインハート〔著〕
西原 史暁〔訳〕

# 目次

# 訳者によるまえがき

嘘には 3 種類の嘘がある。
嘘、くそったれな嘘、そして統計だ。
——ベンジャミン・ディズレーリの言葉とされる

この本は、アレックス・ラインハート (Alex Reinhart) 氏が書いた *Statistics Done Wrong* の全訳である。*Statistics Done Wrong* は科学研究において統計が正しく使われていないことに対して警鐘を鳴らすためにラインハート氏が書いたガイドブックである。このガイドブックは英語の原著がウェブ上 (`http://www.statisticsdonewrong.com/`) で全文公開されている。本和訳はこのウェブ上で公開された文章をもとに翻訳している。また、このウェブ版に加筆をした書籍が 2015 年 3 月にノー・スターチ・プレス (No Starch Press) から出版されるとのことである (ISBN: 978-1-59327-620-1)。

この本では、現在の科学研究において統計が誤用されていることが非常に多く、それがために科学研究の信頼性が揺らいでいることが、様々な例とともに記されている。この本に載っている「ダメな統計」の事例には医学に関するものが多いが、統計が正しく使われていないのは他の科学の分野でも同じである。

だから、科学を研究しようとする人は、ぜひこの本を読んでほしい。分野にもよるが、今の科学で統計を無視して研究することはほとんど不可能だ。科学を研究する人ならば、この本で触れられているような誤った統計分析は誰もが起こしうるのだ。自分自身が統計でつまづかないように、また科学研究のコミュニティーが統計で失敗しないように、科学者は統計についてしっかり理解する必要がある。この本がその理解の一助となれば幸いである。

なお、この本は科学だけでなく他の分野で統計を使う人の参考になるところもあるはずだ。この本で触れられているのは科学研究での統計の問題であるが、同様の問題は、例えばビジネスにおける統計の使用でも発生しうる。このため、科学者以外でもこの本を読む価値があるだろう。

翻訳に当たっては、原文の説明で分かりにくいところを分かりやすくするために、適宜説明の順序を変えたり、原文で用いられている語句の言い換えをしたりするようにした。また、原文で不足している説明を訳注の形で補うようにした。さらに理解を深めるためにいくつか訳者によるコラムを付した。

# 1 はじめに

訳者による概要説明
この章は、本書全体の序論として、科学界に統計上の誤りが広く見られ、それが問題になっていることを説明している。

ダレル・ハフは彼の有名な著作『統計でウソをつく法』[訳注 1]の最後の章で、「医療の専門家らしいところがあるもの」や科学研究室・大学によって公刊されたものは信用する価値があると述べている[訳注 2]。無条件の信用ではないが、メディアやいかがわしい政治家に対する信用よりも確実に信頼が置ける。何しろハフは『統計でウソをつく法』という本の全てを政治やメディアでの人を惑わせる統計的なまやかしで埋めていたぐらいだが、訓練されたプロの科学者による統計について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、理解を追い求めるのだ。

統計データの分析は科学にとっての基礎である。お気に入りの医学誌の中からランダムに 1 ページを開けば、統計――*t* 検定、*p*値、比例ハザードモデル、リスク比、ロジスティック回帰、最小二乗適合、信頼区間――に圧倒されるだろう。統計学者は複雑なデータセットのほとんどに対して秩序と意味を見いだすための巨大な力を持つ手法を科学者に提供し、科学者は大喜びでこうした手法を受け入れてきた。

[訳注 1] ダレル・ハフ (Darrell Huff, 1913-2001) はアメリカの著述家である。統計の専門家というわけではなかったが、彼が書いた『統計でウソをつく法』 (*How to Lie with Statistics*) は英語版だけで 50 万冊以上売れた [文献 59]。

[訳注 2] ここの部分だけ見ていると、ハフが医療専門家などを信用すべきだと言っているようにも見えるが、ハフは別に医療専門家などを信用すべきだと言っていたわけではない。『統計でウソをつく法』ではむしろ有名大学からのものだと述べるはったりから信用しないように気をつけようと言っている。

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、科学の学部課程の多くで、統計の訓練を全く求めていない。

1980 年代以降、研究者は評判の良い査読[訳注 3]つきの科学文献における多数の統計的な誤謬と誤解を説明し、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示している。多くの研究で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いる。問題はごまかしが行われていることではない。問題は貧弱な統計教育だ。公刊された研究上の発見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［文献 33］。

以下に記されるのは、科学の名の下に日常的に行われるとんでもない統計の誤りのリストである。以下では、統計の方法についての知識がないものとして話を進める。多くの科学者は正式な統計の訓練を受けていないからだ。そして、注意しておこう。こうした誤りを一度学んだら、あちこちでその誤りを見ることになるだろう。驚かないでほしい。このことは、現代科学の全てを否定し、瀉血とヒル[訳注 4]に戻すための根拠とはならない。これは、我々が必要としている科学を改善するための要望なのだ。

［訳注 3］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、掲載する前に届いた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始めて論文として学術誌に掲載され、公刊される。

［訳注 4］かつての医学では、血を抜くこと（＝瀉血）が病気の治療になると考えていた。そして、血を抜くために吸血性のヒルという虫が用いられた。今では瀉血は科学的な根拠がないと否定されている。

# 2 データ分析入門

訳者による概要説明
この章は、統計分析でよく用いられる $p$ 値という概念について説明している。この概念は本文で触れられているように非常に誤解されやすい概念である。きちんとした統計分析を行うにはこの概念をしっかり理解しなくてはならない。

実験科学の多くは、つまるところ、違いを測定するということに行き着く。例えば、ある薬は他のものよりうまく働くか、ある種の遺伝子を持つ細胞は他の種類の遺伝子を持つ細胞より酵素をたくさん合成するか、ある種の信号処理アルゴリズムは他のものよりパルサー[訳注 5]をよく検出できるか、ある触媒は化学反応をより効果的に加速するかといったたぐいの問題だ。

となると、統計の多くは、こういった差について判断をすることに行き着くことになる。まずは「統計的有意差」を話題にしよう。偶然以外の何かのせいだと言えるほど測定と測定の間の差が本当に大きいかについて判断する方法を統計学者が工夫してきたからだ。

あなたがかぜ薬を試験しているとしよう。あなたの新薬を使うと、かぜの症状が続く期間が 1 日短くなると期待されている。このことを証明するために、かぜをひいた患者を 20 人見つけ、その半数に新薬を、残りの半数に偽薬[訳注 6]与えたとしよう。そして、かぜの長さを調べ、薬のあるなしによってかぜの長さの平均がどうなるのか分かったとしよう。

[訳注 5] パルサー (pulser) とは、短い周期で電波や X 線を発する天体のことである。

[訳注 6] 偽薬 (placebo) とは、見た目こそ普通の薬のようだが、実際には薬としての効果が全くないもののことである。なお、偽薬を患者に与える場合、普通は薬としての効果がないとは言わないでおく。

だけれども、かぜは全てが同じというものではない。平均的なかぜは 1 週間続くかもしれないが、数日しか続かないかぜもあるし、2 週間かそれ以上続いて家の中にあるティッシュペーパーを全て使い果たすほどのかぜだってある。本物の薬を投与された方の 10 人の患者が、2 週間のかぜをひく不幸なタイプだった場合、新薬はかえって状況を悪化させると間違った結論を出してしまうかもしれない。どうすれば、不幸な患者がいると示すのではなく、あなたの薬が機能することを示すと判断できるのだろうか。

## *p* 値の力

統計がその答えを示してくれる。もし、典型的なかぜの症例の分布——短いかぜ、長いかぜ、平均的なかぜのそれぞれにどれだけの患者がかかるかという大まかな話——を知っていれば、かぜ患者を無作為に選んだ標本で、平均より短いかぜ、平均より長いかぜ、ちょうど平均のかぜがどれだけありそうかを判断できる。統計的検定を行うことで、「もし私の薬が全く効果がなかったとしたら、私が観察したようなデータを観察する確率はどれほどか」といった質問に答えることができる。

これは、ちょっとややこしいから、もう一回読んでほしい。

直感的には、このことがどう働くかを理解できる。1 人に対してしか薬を試していないとき、患者の約半分は平均より短いかぜになるのだから、その人が平均より短いかぜになったとしても何も驚くことはない。1,000 万人の患者に対して薬を試したとき、その薬が機能していない場合、全員が平均より短いかぜになることはものすごくありえないことだ。

科学者が用いる一般的な統計的検定では、*p* 値という数値が出てくる。この数値は、上に述べたことを数量の形で表したもので、以下がその定義だ。

> *p* 値は、効果がないか、差異がないという仮定（帰無仮説）のもとで、実際に観測された結果と同じか、それよりも極端な結果が出る確率として定義される［文献 24］。

だから、100 人の患者に薬を与えて、これらの患者のかぜが平均して 1 日短いことが分かった場合、この結果に対する $p$ 値は、薬が全然働かなかったときに 100 人の患者がたまたま 1 日短いかぜをひいていた確率のことなのだ[訳注 7]。明らかに、この $p$ 値は効果の大きさ——かぜが 4 日短いのはかぜが 1 日短いことよりもありえそうにない——と薬物治療の調査を行った患者の数に依存する。

これは、理解するにはややこしい概念だ。$p$ 値というのは、正しさを測定するものでなければ、違いがどれだけ重大かを測定するものでもない。$p$ 値とは、グループ間で本質的な違いがないにもかかわらず、違いがあることを示唆するデータが得られたときに、どれだけ驚くべきかを示す値なのだ。より大きな差異があったり、より多くのデータによって支えられたりしたものは、より驚くべきことを示唆し、より小さい $p$ 値を示唆する。

このことを「本当に違いはあるのか」という問いに対する答えに翻訳することは簡単ではない。ほとんどの科学者は、単純でおおざっぱなやり方を使っている。もし $p$ が 0.05 より小さければ、薬が本当は働いていない場合にこうしたデータを得る確率は 5% しかないわけだから、薬と偽薬の間の差が「有意である」と呼ぶのである[訳注 8]。もし、$p$ が 0.05 より大きければ、差は有意でないと呼ぶ。

しかし、限界がある。$p$ 値は驚きを測定するもので、効果のサイズを測定するものではない。「この薬は 4 倍長生きさせる」といった極めて大きい効果を測定するか、ごく小さな効果だが非常に確実な効果を測定することで、きわめて小さい $p$ 値を得ることができる。統計的に有意であることは、結果が実際に意味があるものであることを意味しない。

---

[訳注 7] 原文のここでの $p$ 値の説明はあまり正確ではない。より正確に言えば、「薬が全然働かなかったときに 100 人の患者が平均してたまたま 1 日以上短いかぜをひいていた確率」になる。

[訳注 8] ここの例で 0.05 より大きいか小さいかが有意であるかどうかの基準になっている。こうした基準は、有意水準 (significance level) と呼ばれる。科学研究においては、慣例的に 0.05 を有意水準とすることが多い。しかし、0.05 を有意水準とすることは単なる慣習に過ぎず、この数値を有意水準として選ぶ客観的な根拠があるわけではない。

同様に、統計的に有意でないことも解釈しにくい。完璧にすばらしい薬があったとしても、それを 10 人にしか試さなかったとしたら、患者に対する本当の改善と単なる幸運との違いを見分けることは困難だろう。あるいは、何千人もの人に試すことができたとしても、その薬が 3 分間しかかぜを縮めないとしたら、差を検出することは単純に不可能であろう。統計的に有意な差がないことは、差が全然ないことを意味しないのだ。

仮説が本当かどうかを判断する数学的な手段はない。仮説がデータと矛盾していないかを見ることしかできない。そして、データが足りなかったり、はっきりしなかったら、結論は明確なものにはならない。

だが、我々はそれでやめるわけにはいかないのだ。

### ❖ 訳者コラム：$p$ 値をどう報告するか ❖

本文では $p$ が 0.05 より小さいかどうかで、有意であるかどうかを判断するとしている。このことを踏まえると、学術論文で $p$ 値を報告するときは $p < 0.05$ のように不等式の形で書けばよいと考える人もいるかもしれない。実際、多くの学術論文で $p < 0.05$ のように不等式だけが書かれている。

しかし、このような書き方は望ましくない。例えば $p = 0.0498$ と $p = 0.0113$ では値が全く異なるのに、両方とも $p < 0.05$ となってしまうからだ。このようなことを防ぐために、$p = 0.0323$ のように、$p$ の具体的な値を書くべきである。この値は統計ソフトを使えば簡単に計算することができる。

なお、本書では $p = 0.05$ のように小数点の前にゼロを付しているが、このゼロを省略するスタイルもある。こうしたスタイルでは、$p = 0.341$ でなくて、$p = .341$ と表記する。$p$ 値は 1 未満の値で小数点の前に 0 以外の数字が来ることはないので、省略しても問題はないのだ。

# 3 検定力と検定力の足りない統計

訳者による概要説明
この章では、検定における重要な概念の１つである「検定力」について触れた後、検定力が足りないために適切な結果が得られないことがあることを説明している。

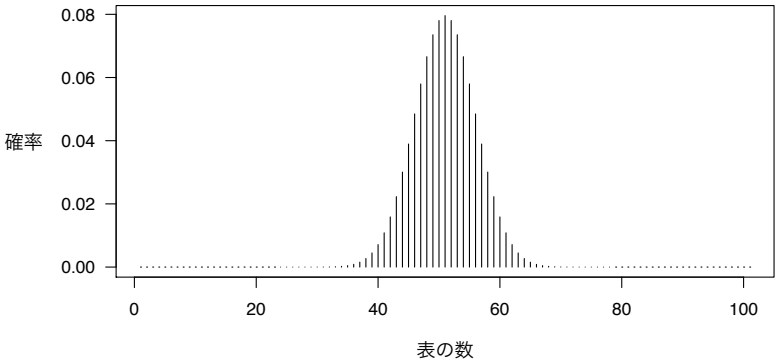
単に十分なデータを取らないだけで、実際に存在する効果を見つけられない可能性があることを見てきた。ほとんどの場合、これは問題となる。うまくいきそうな薬を見つけられなかったり、重大な副作用に気づかなかったりするかもしれない。データをどれだけ集めれば良いかということは、どうすれば分かるだろうか。

統計学者は「検定力」(statistical power)[訳注 9]という形で答えを用意している。ある研究における検定力とは、単なる幸運からある程度の大きさのある効果を区別する可能性のことを指す。薬から得られる大きな利益は簡単に検出できるだろうが、微妙な違いを検出することはずっと可能性が低い。簡単な例を見てみよう。

ギャンブラーが、相手が不正なコインを持っていると確信しているとしよう。このコインは、表と裏が出る割合が半々ではなく、出る割合が違っている。そして、相手は非常に退屈なコイン投げゲームで不正をするためにこのコインを使っている。このことをどう確かめれば良いだろうか？

単にコインを 100 回投げて、表が出た回数を数えるのではだめだ。たとえ、完全に公正なコインであったとしても、いつも 50 回表が出るわけではない。

[訳注 9] 検定力は、検出力や統計力と呼ばれることもある。

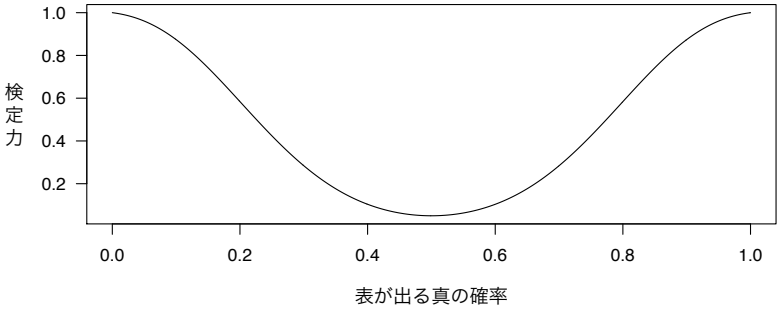



これは、100 回コインを投げたときに、表が何回出るかという可能性を示したものである。

50 回表が出る可能性が一番高いことが分かるが、45 回表が出たり、57 回表が出たりする可能性もかなり高い。だから、57 回表が出たとしたら、コインは不正なものだったかもしれないが、単に運がよかっただけかもしれない。

数学的に解いてみよう。例えば、科学者がよくするように、$p$ 値が 0.05 以下になるところを探そう。つまり、10 回または 100 回の試行の後に、表の出た数を合算し、表が半分で裏が半分になるだろうという期待からのずれを求める。公正なコインではそれ以上のずれが起きる可能性が 5% しかなければ、コインが不正なものだということにしよう。そうでなければ、何も結論づけることはできない。コインは公正かもしれないし、少し不正かもしれない。判断が付けられないのだ。

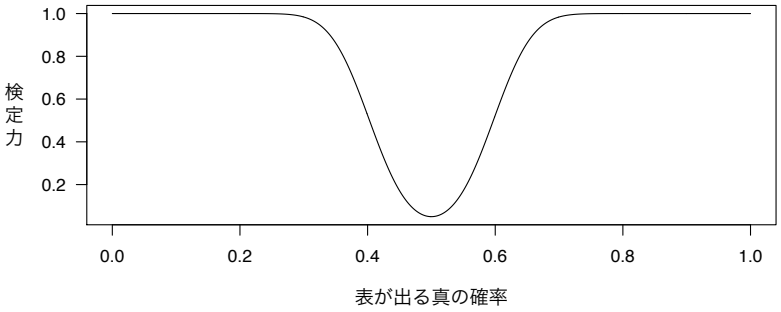
さて、もし 10 回コインを投げて、上記の基準を適用したとしたら、どうなるだろうか。




これは検定力曲線と呼ばれるものだ。横軸にそって、コインの表が出る真の確率ごとに異なった可能性が示されている。コインの表が出る真の確率は不正さの度合いに対応している。縦軸は、10 回投げたあと、その結果に対する $p$ 値をもとに、コインに不正があると結論づける確率だ[訳注 10]。

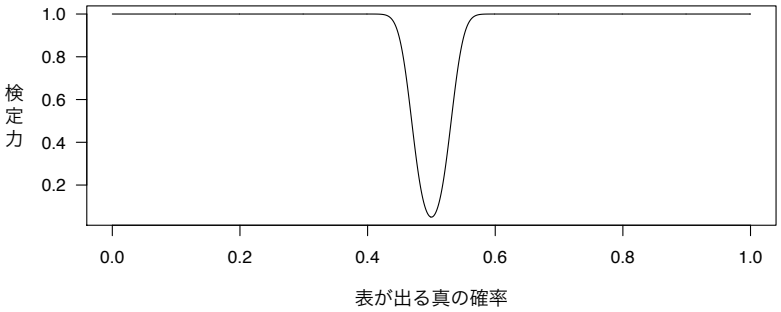
コインが不正なもので 60% の確率で表が出るものであるとき、10 回投げた場合、不正があると結論づけられる確率は 20% しかない。データが少なすぎて、無作為な変動から不正を区別することができないのだ。不正に必ず気づくためには、コインが信じられないほど偏ってなくてはならない。

しかし、もし 100 回コインを投げたとしたら？




[訳注 10] 10 回投げて得られた結果に対して、二項検定と呼ばれる検定を行い、$p$ 値を求める。ここで、$p < 0.05$ ならば、コインに不正があると結論づけることになる。

あるいは 1,000 回ならば？




1,000 回投げれば、コインに不正があって 60% 表が出るということを簡単に判断することができる。公正なコインを 1,000 回投げて、600 回以上表が出ることは、ほとんどありえないからだ。

## 足りない検定力

こうしたことを聞けば、検定力の計算が、医学に関する試験において必要不可欠であると考えるかもしれない。科学者は、新しい薬が生存率を 10% 以上改善するかを試験するためにどれだけの患者が必要なのかについて、知りたいかもしれない。検定力を手っ取り早く計算することでその答えが得られるだろう。一般に、科学者は検定力が 0.8 以上であれば満足する。これは、実際の影響があると結論づける確率が 80% であることに対応する。

しかし、この計算をする科学者はほとんどいないし、学術誌の記事で検定力を報告する記事もほとんどない。

同じ条件の下で、2 つの異なった薬を与える治療を試すことを考えてみよう。どちらの薬がより安全なのか知りたいところなのだが、あいにくどちらの薬も副作用がまれであるとする。各々の薬は 100 人の患者に対して試験できるものの、各グループで深刻な副作用が生じるのはほんの少ししかないものとする。

明らかに、副作用の割合を比較するために十分なデータがない。もし片方のグループで 4 人に深刻な副作用が生じ、もう片方のグループで 3 人に深刻な副作用が生じたとしたら、その違いが薬のせいかどうか判断できないのだ。

不幸なことに、十分なデータがないためにきわめて大きな違い以外は検出できないということに言及せずに、「有害な影響に関して、グループの間に統計的に有意な差がない」という結論をしている試験はたくさんある［文献 62］。そして、そのために、片方はもう片方よりずっと危険かもしれないのに、医師が 2 つの薬は同じぐらい安全だと誤って考えてしまうのだ。

これは、弱い効果しか持たない薬に関する問題でしかないと思うかもしれない。しかし、違うのだ。1975 年から 1990 年までの間に権威ある医学誌に公刊された研究から抽出されたある標本では、ランダム化比較試験［訳注 11］の 27% が否定的な結果［訳注 12］を示していたのだが、そのうち 64% が、グループの間の主要評価項目の 50% の違いを明らかにするのに十分なデータを使っていなかったのだ。50%! たとえ、ある薬が他の薬に比べて症状を 50% 減らすとしても、その薬がより効果的だと結論づけるために十分なデータがないのだ。そして、否定的な結果を示した試験の 84% が、25% の差を発見する検定力がなかったのだ［文献 4,11,16,46］。

神経科学では、問題はもっと悲惨だ。ある特定の効果を調べているたくさんの神経科学の論文から集められたデータを合算して、効果量について強い推定量が出てきたとしよう。合算対象となっている個々の研究のうち、中央値にあたる研究がこの効果を検出できる可能性は 20% しかない。たくさん

［訳注 11］ランダム化比較試験 (randomized controlled trial; RCT) とは、主に医学の分野で使われる科学実験の実験デザインの 1 つである。このデザインでは、効果を調べたい治療法を施すグループ（実験群）とそれと比較するためのグループ（対照群）を設けて、効果がどれだけあるかを調べる。また、被験者を偏りなく選び出した上で、偏りなく実験群と対照群に割り当てを行うこともこの実験デザインの重要な特徴である。

［訳注 12］2 つのグループの違いを比べようと実験を行った場合、2 つのグループの間に明確な差が見いだせない場合がある。否定的な結果 (negative result) とは、そういった場合のことを指す。

の研究を合算してはじめて効果を発見できるのだ。同様な問題は、実験動物を使う神経科学の研究でも起きており、大きな倫理的問題を引き起こしている。もし、個々の研究の検定力が足りなければ、本当の効果は多くの動物を使ったたくさんの研究が終了して解析されてからでないと本当の効果は発見されないだろう。最初にちゃんと研究が行われるよりずっと多くの実験動物を使うのだ[訳注 13][文献 12]。

これは科学者がグループ間で有意な差がないと述べているとき、科学者が嘘をついていると言うものではない。実際の違いがないことを意味すると誤解してしまっているだけだ。違いがあるかもしれないが、研究の規模が小さすぎて違いに気づくことができないのだ。

日常的な例を考えてみよう。

## 赤信号での誤った方向転換

1970 年代、アメリカの多くの地域で、赤信号で車が右折することが許されるようになった。それに先立つ長い間、道路の設計者と土木技師は赤信号での右折を許すと衝突や歩行者の死亡が増えるので危険であると主張してきた。しかし、1973 年の石油危機とその影響により、通勤する人が赤信号を待って燃料を無駄にするしないように、赤信号で右折することを許すべきだと政治家が考えるようになった。

この変化が安全に対して与える影響を考察するためにいくつかの研究が

[訳注 13] 例えば実験動物を 200 匹使えば、十分な検定力が得られて、効果があると主張できるとしよう。この場合、誰かが最初に 200 匹を使って効果があると論文に書けば、犠牲になる実験動物は 200 匹しかいないことになる。しかし、実験する人がそろいもそろって 50 匹ずつしか使わなかった場合はどうなるだろうか。1 人目は 50 匹しか使わず効果があると主張することに失敗する。2 人目がまた別の 50 匹を使って効果があると主張することに失敗し、3 人目がまた別の 50 匹を使って……ということが延々と続くことになる。50 匹ずつ使った人が合わせて 10 人いたとすればその時点で 500 匹も実験動物が犠牲になる。誰かがこれら 10 人分のデータを集めて分析することがあれば、500 匹分のデータがあるので効果があると主張できる。しかし、誰も集めなければ、500 匹もの犠牲が生じた上に、何も知見が得られないということになってしまう。

行われた。例えば、バージニアの高速道路および交通部門のコンサルタントは、赤信号でも右折が許されるようになった 20 箇所の交差点で、変化前と変化後の違いを調べる研究をした。変化前は、これらの交差点で事故が 308 回あった。変化後は、同様の長さの期間で事故が 337 回あった。しかし、この差は統計的に有意でなかったため、コンサルタントは安全に対する影響がないとの結論を出した。

これに続くいくつかの研究も同じような結果だった。すなわち、衝突回数は少し増加するが、その増加量が統計的に有意であると結論するには十分なデータがないということだ。ある報告は以下のような結論を述べている。

> （赤信号での右折の）採用以降、右折が関わる歩行者事故が増加したと疑う理由はない……

このデータに基づいて、さらに多くの市や州が赤信号での右折を許すようになった。もちろん、問題はこれらの研究が検定力が足りないことである。車にひかれる歩行者が増え、衝突に巻き込まれる車も増えたのだが、このことを確信をもって示すための十分なデータを誰も集めることができなかった。数年後、衝突と歩行者の事故が有意に増加している（時には 100% 近くの増加もあった）という明確な結果がもたらされるまでは［文献 29,52］。検定力の足りなかった研究に対する誤った解釈が命を奪ったのである。

# 4 擬似反復：データを賢く選べ

訳者による概要説明
この章では、データ数を増やすために行われる「擬似反復」の問題について説明している。

多くの研究では、反復をすることによって、より多くのデータを集めようと努力している。追加の患者や標本に対して測定を繰り返すことで、数値についてよりはっきりさせることができ、パッと見ただけでは明らかにはならないような目立たない関係を発見することができる。検定力を高めたり小さな違いを見つけたりする時の追加データの価値についてはすでに見てきた。だが、実際のところ何が反復として扱われるのだろうか。

医学の例に戻ってみよう。100 人の患者グループが 2 つあり、それぞれに異なった薬を投与したとき、どちらの薬が血圧をより下げるのかを明らかにしたいものとする。各グループに対し、効果が出るように 1 ヶ月間薬を飲ませ、その後各グループについて 10 日間毎日血圧を測る。そうすると、患者ごとに 10 のデータ点があり、グループごとに 1,000 個のデータ点があることになる。

すばらしい！　1,000 個のデータ点というのはとっても多い。片方のグループがもう片方のグループに比べて血圧が低いということがかなり簡単に確かめられる。統計的有意性を計算すれば、とても簡単に有意な結果が得られる。

でも待ってほしい。1 人の患者について 10 回血圧を測れば、10 個のよく似た結果が得られると予想される。もし、ある患者が遺伝的に低血圧の傾向にあれば、その遺伝的特徴を 10 回測っていることになる。100 人の測定を繰り返す代わりに、1,000 人の別々の患者からデータを集めたとしたら、グ

ループ間の違いは薬に起因するのであって、遺伝的特徴や運によるものではないと、より自信を持って言えただろう。ここで標本サイズが統計的に有意な結果と高い検定力を与えるほど大きいと主張したとしても、この主張は正当なものではないのだ。

この問題は擬似反復 (pseudoreplication) として知られていて、きわめてありふれたものだ［文献 41］。ある培養物からの細胞を調べたあとに、同じ培養物からより多くの細胞を取り出して調べることによって、生物学者は結果を「反復する」かもしれない。たった 2 匹のラットから何百ものニューロンを調べたので標本サイズが大きいと誤って主張するなど、神経科学者は同じ動物からの複数のニューロンを調べるかもしれない。

統計学的な言い方に従えば、擬似反復は個々の観察が互いに強く依存している［訳注 14］ときに起きる。ある患者の血圧の測定結果はその患者の前日の血圧と強く関連しているし、ある場所の土壌組成の測定結果は 5 フィート（約 152 センチメートル）先の場所の測定結果と強く相関しているだろう。統計分析を行う際に、こうした依存を説明する方法はいくつかある。

1. 独立していないデータ点の平均をとる。例えば、ある個人から測定された血圧の平均をとる。だけれども、これは完璧な方法ではない。もしある患者について他の患者よりたくさん測定を行ったとしても、そのことは平均の数値に反映されない。より多くの測定が行われるほど、より信頼度が高くなる方法が必要だ。
2. 独立していないデータ点を別個に分析する。1 人から 1 つのデータ点を取り出す形で、各患者の 5 日目の血圧を分析することができるだろう。しかし、注意する必要がある。なぜならば、こうしたことを毎日行えば、次の章で議論することになる多重比較の問題を引き起こすからだ。

［訳注 14］「依存している」ということは、すなわち独立していないということを指す。

3. 階層モデル[訳注 15]やランダム効果モデル[訳注 16]のように、独立していないことを説明する統計的モデルを用いる。

各手法が適する状況は異なるので、データを分析する前に各手法を検討することが重要だ。擬似反復は、被験者に対する追加的な情報をほとんど提供しないのにもかかわらず、有意差を出すことを簡単にする。標本を再び調べることを通じてわざと標本サイズを大きく見せることをしないように研究者は注意しなくてはならない。

[訳注 15] 階層モデル (hierarchical model) とは、あるものが別のものに含まれているという階層関係を説明に入れて組み立てる統計的モデルである。例えば、関東の家庭と近畿の家庭とで年間の靴の購入数に違いがあるかを調べるとしよう。そして、関東から 5 都市（東京、横浜、千葉、水戸、宇都宮）、近畿から 5 都市（大阪、京都、神戸、奈良、大津）を選び、それぞれの都市から 10 の家庭を選ぶものとしよう。靴の購入数は、関東か近畿かで変わるかもしれないし、都市によって変わるかもしれない。だから、モデルを組み立てるときは、関東地方か近畿地方かということを組み込むほか、どの都市かということも組み込まなくてはならない。しかし、このとき、地方と都市とを独立したものとして考えてはならない。都市が決まればどの地方にあるかを確定できるので、両者は独立していないのである。もし独立しているとしたら、地方と都市は無関係ということになって、関東にある京都や近畿にある千葉というものが設定できるという変な状態になってしまう。だから、「都市が地方に含まれている」という階層関係をモデルに組み込んで統計分析を行わなくてはならない。このように分析することでより適切な分析が可能になるのだ。

[訳注 16] ランダム効果 (random effect) と対になるものとして固定効果 (fixed effect) というものがある。ランダム効果と固定効果がどういうものなのかについての説明は *Modern Statistics for the Life Sciences* という本 [文献 28] の第 12 章が分かりやすい。ランダム効果なのか固定効果なのかを正しく決めないと、正確な統計分析はできなくなってしまう。本来はランダム効果として扱うべきなのに、固定効果のように扱ってしまって分析がおかしくなることはしばしば存在する [文献 17]。

❖ 訳者コラム：偽陽性、偽陰性、第一種の誤り、第二種の誤り ❖

後に続く章では「偽陽性」(false positive) や「偽陰性」(false negative) というものが扱われるので、先にこれらの用語について説明しておこう。

陽性や陰性というのは医学検査でしばしば用いられる用語だ。例えばある病気に感染しているかを検査した時に、感染しているという結果が出た場合を陽性と呼び、出なかった場合を陰性と呼ぶ。ここで注意しなくてはならないのは、検査が 100% 正しいわけではないということである。つまり、本当は感染していないのに、感染しているという結果が出ることがある。これを偽陽性と呼ぶ。また、本当は感染しているのに、感染しているという結果が出ないことがある。こちらは偽陰性と呼ぶ。

偽陽性と偽陰性という概念は医学検査以外でも用いられることがあり、統計の検定についてもこれらの概念が用いられる。検定においては、本当は帰無仮説が正しいのに有意であるという結果が出てしまうことを偽陽性とする。なお、検定における偽陽性のことを第一種の誤り (type I error) と呼ぶこともある。これに対して、本当は帰無仮説が正しくないのに有意であるという結果が出ないことを偽陰性とする。検定における偽陰性は第二種の誤り (type II error) と呼ぶこともある。

どんな検定においても、第一種の誤りと第二種の誤りを完全に消し去ることはできない。逆に言うと、検定を行う際はいつでも第一種の誤りや第二種の誤りを起こす可能性があるということに注意しなくてはならない。

また、第一種の誤りをする可能性と第二種の誤りをする可能性はトレードオフである。つまり、第一種の誤りをできるだけしないようにした場合、第二種の誤りをする可能性が増える。逆に、第二種の誤りをできるだけしないようにした場合、第一種の誤りをする可能性が増える。

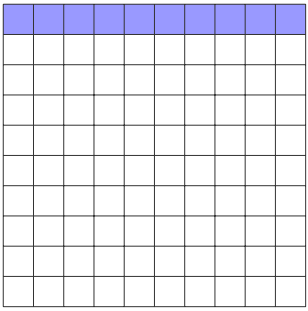
# 5 $p$ 値と基準率の誤り

訳者による概要説明
この章では、何度も比較を行うことの問題点について扱っている。何度も比較を行えば、本当は存在しない現象が存在するかのように判断されてしまう可能性がある。

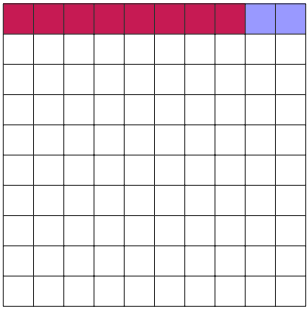
$p$ 値が解釈しにくいことについてはすでに見てきた。統計的に有意でない結果が得られたからといって、違いがないことを意味するわけではない。では、有意な結果が得られた場合はどうだろうか。

例を 1 つ見てみよう。ガンを治す見込みがある薬を 100 種類試験しているとする。これらの薬のうち、実際には 10 種類しか効かないのだが、どれが効くのかは分からない。よって、効く薬を見つけるために実験をしなくてはならない。実験においては、薬に有意な利点があることを示すために、偽薬に対して $p < 0.05$ となる薬を探すことになる。

これを図示してみよう。図の 1 マスが 1 種類の薬を表している。青いマスは、効く薬を示す。

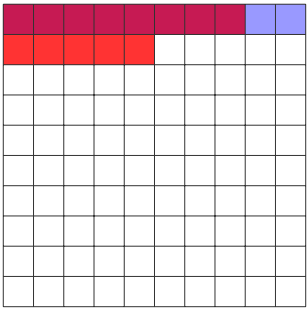


既に見てきたように、ほとんどの試験は、全ての良い薬を完全に発見できるわけではない。検定力が 0.8 であると仮定しよう。このとき、10 種類の良い薬のうち、紫色で示された 8 種類を正確に検出することができるだろう。



そして、90 種類の効果のない薬のうち、5 種類が有意な効果があると結論づけることになるだろう。なぜか。$p$ 値は効果がないという仮定の下で計算されるということを思い出してほしい。だから、$p = 0.05$ というのは、効果がない薬が効いていると誤って結論づける可能性が 5% あるということを示しているのだ[訳注 17]。

よって、実験を行うと、効く薬が 13 種類あると結論づけることになる。良い薬が 8 種類と、赤で示された誤って含めてしまった薬が 5 種類だ。



[訳注 17] 厳密に言えば、90 種類の薬の 5% なので、4.5 種類ということになる。ただ、種類の数が整数個にならないと分かりにくいので、切り上げて 5 種類ということにしたのであろう。

どの「効く」薬についても、本当に有効である可能性はたった 62% である。もし、100 種類の中から無作為に薬を 1 種類選んで試験を行い、$p < 0.05$ という統計的に有意な利点を発見したとしても、その薬が実際に有効である可能性は 62% しかないのである。統計学の用語で言うと、偽発見率（本当は偽陽性であるのに統計的に有意な結果が出る割合）が 38% になるということなのである。

ここでは、有効なガン治療薬の**基準率**[訳注 18]がとても低い——100 種類の臨床試験薬のうち実際に効くのは 10% だけである——ので、試験される薬は多くが効かず、偽陽性に遭遇する機会が多い。もし、完全に効果がない薬をトラック 1 台分あるという不幸に襲われれば、基準率が 0% ということになり、統計的に有意な結果が真である可能性は 0% になる。にもかかわらず、トラックの中の薬のうち 5% について、$p < 0.05$ という結果が得られる。

「$p$ 値はエラーがありえないことを示す兆候だ」と言う人を見たことがしばしばあるだろう。そう述べる人は、$p = 0.0001$ という結果を得て、「統計的な偶然としてこの結果が出てくるのは 1 万回に 1 度しかない」[訳注 19]と言う。それは違う。これは基準率を無視している。そして、このことは**基準率の誤り**と呼ばれる。$p$ 値がどう定義されるか思い出してみよう。

> $p$ 値は、効果がないか、差異がないという仮定（帰無仮説）のもとで、実際に観測された結果と同じか、それよりも極端な結果が出る確率として定義される。

$p$ 値は薬が**有効でない**という仮定の下で計算され、得られたデータと同じ

[訳注 18] 調査対象となっているもののうち、真に有効であるものの割合を基準率 (base rate) と呼ぶ。ここのガン治療薬の例で言えば、100 種類の薬のうち、真に有効なのは 10 種類なので、$10 \div 100 = 0.1 = 10\%$ が基準率となる。

[訳注 19] $p$ 値の意味について誤解している人は、しばしば「統計的な偶然としてこの結果が出てくるのは 1 万回に 1 度しかない」と述べるのに加えて、「だから、有効である確率は 1 万回に 9,999 回、つまり 99.99% だ」と述べることがある。もちろん、こう考えるのは誤りである。

か、さらに極端なデータが得られる確率について教えてくれる。薬が有効である確率については教えてくれないのである。

「おそらく正しいだろう」と述べるために $p$ 値を使う人がいたら、このことを思い出そう。そうした人の研究が誤っている確率は、ほとんど確実にぐんと高い。開発初期段階の薬の試験（こうした薬は試験を切り抜けることがほとんどない）のように、ほとんどの検定された仮説が偽となる分野においては、$p < 0.05$ となる「統計的に有意な」結果のほとんどが実際にはまぐれあたりである可能性が高い。

良い例が、医療上の診断検査だ。

## 医療検査における基準率の誤り

乳ガンのスクリーニング[訳注 20]にマンモグラフィー[訳注 21]を使うことについて、論争がある。不必要な生体組織検査・手術・化学療法といった偽陽性の結果でもたらされる危険の方が、ガンの早期発見の利益を上回ると主張する人もいる。これは統計の問題だ。このことについて数値的に評価してみよう。

マンモグラフィーを受けた女性のうち、0.8% が乳ガンであるとしよう。乳ガンの女性のうち、マンモグラフィーで乳ガンが正確に発見できるのは、90% である。（90% というのはこの検査の検定力に相当する。そこにガンがあると分からないのであれば、どれだけのガンが見逃されているか判断しがたいという意味で、これは推定量に過ぎない。）ただし、全く乳ガンにかかっていない女性のうち、約 7% がマンモグラフィーで陽性が出て、さらなる生体組織検査などの検査が必要になる。もし、マンモグラフィーで陽性が出た場合、乳ガンにかかっている確率はどれぐらいだろうか？

[訳注 20] スクリーニングとは、病気の疑いがある人を選別することである。

[訳注 21] マンモグラフィーとは触診では分からないような小さな乳ガンを発見するために、乳房に対して行われる X 線検査のことである。

検査対象者が男性である可能性[原注 1][訳注 22]を無視すれば、この答えは 9% になる［文献 38］。

ガンを患っていない女性の 7% にしか偽陽性にならない検査（これは $p < 0.07$ である検定に類似する）であるにもかかわらず、陽性の結果が出た場合の 91% が偽陽性なのである。

これはどう算出されたのだろうか。ガンの治療薬の例と同じ方法によって算出している。マンモグラフィーを受けることを選んだ女性の中から無作為に 1000 人選んだとしよう。そのうち、8 人（0.8%）が乳ガンにかかっている。マンモグラフィーは、乳ガンの場合の 90% を正確に発見するので、8 人中 7 人の女性についてガンが発見されることになる。しかしながら、992 人の乳ガンではない女性がおり、そのうち 7% がマンモグラフィーで偽陽性の結果を得る。つまり、70 人の女性[訳注 23]が誤ってガンであるとされてしまうのである。

合計すると、77 人の女性がマンモグラフィーで陽性となり、そのうち 7 人が実際に乳ガンであることになる。マンモグラフィーで陽性だった女性のうち、9% しか乳ガンにかかかっていないのである。

もし統計学の学生や科学の方法論の講師にこのような質問をしたら、3 分の 1 以上が間違える［文献 38］。もし医者に聞いたら 3 分の 2 が間違える［文献 10］。彼らは $p < 0.05$ という結果は 95% の確率でその結果が正しいと

［原注 1］興味深いことに、男性であることは乳ガンにかかる可能性を排除しない。男性であることは、乳ガンになる可能性を非常に低くするにすぎない。

［訳注 22］乳ガンは女性に多い病であるが、男性でもまれにこの病にかかることがある。日本乳癌学会が 2014 年に出した『全国乳がん患者登録調査報告：2011 年次症例』(`http://www.jbcs.gr.jp/people/nenjihoukoku/2011nenji.pdf`) によれば、2011 年の日本における乳ガン発症数として、女性の 48,262 症例と男性の 219 症例が報告されている。なお、これは 2011 年に「報告」された数なので、同年に日本で実際に乳ガンを発症した人を完全に網羅した数ではない。また同一患者が両側の乳房でガンになった場合は 2 症例と数えられている。

［訳注 23］厳密に言えば、992 人の 7% なので、69.44 人となる。分かりやすくするために、キリの良い 70 人にしたのであろう。

いうことを意味すると間違った結論を下すのである。しかし、今までの例から分かるように、陽性の結果が真となる可能性は、検定された仮説が真である比率に依存する。幸運なことに、いつもわずかな比率の女性しか乳ガンにかかっていないのだ。

統計の入門の教科書を調べてみれば、同種の誤りがたびたび見つかるだろう。$p$ 値は直感に反するものであり、基準率の誤りはどこにでもあるのだ。

## 基準率の誤りに対して武器をとれ

基準率の誤りをおかすために、先進的なガン研究や早期のガンのスクリーニングを行う必要はない。社会に関する研究を行っている場合はどうだろうか。アメリカ人が自衛のために銃をどれだけの頻度で使うのか調査したいとしよう。銃規制に関する議論は、結局のところ、自衛の権利が中心となっている。このため、銃が防衛のために広く使われているかどうかについて、そして自衛のための銃の使用が殺人などの否定的な面がある銃の使用を上回っているかどうかについて、確認することが重要である。

このデータを得る方法として、調査を通じてデータを手に入れるということがあるだろう。アメリカ人の代表的標本に対して、銃を持っているかどうか、持っているとしたら盗みなどを目的とした住居侵入から家を守ったり路上強盗から身を防いだりするために銃を使ったことがあるかを問うことができるだろう。こうして得られた数値を、法執行機関の統計[訳注 24]から得られる殺人での銃使用の数値と比べることができよう。そして、利点が否定的な面を上回っているかどうかについて、データに基づいて判断することができるだろう。

このような調査は実際に行われたことがあり、興味深い結果を残している。1992 年に行われたある電話での調査では、アメリカの民間人が自衛の

[訳注 24] 原文は “law enforcement statistics” と書かれている。アメリカでは、警察 (police) 以外にも犯罪捜査に当たる公的組織が多数存在し、それらをまとめて法執行機関と呼ぶ。日本の感覚からすれば、「警察統計」という意味になる。

ために銃を用いたことが毎年 250 万回に達すると推定している。すなわち、アメリカの成人の 1%[訳注 25]が小火器で身を守ったということだ。さて、そのうち 34% が盗みなどの犯罪目的の住居侵入に対してのものである。よって、84 万 5 千件の住居侵入が銃の所有者によって防がれたことになる。しかし、1992 年には誰かが家にいるときに行われた犯罪目的の住居侵入は 130 万件しか起きていなかった。そのうち、3 分の 2 は家の所有者が眠っていたときに発生し、侵入者が去った後に発覚したものであった。つまり、家の所有者が家にいて起きた状態で侵入者と対面した住居侵入は 43 万件あり、そのうち我々が信じ込まされているように 84 万 5 千件が銃を持ち歩く住人によって防がれたのだ［文献 30］。

あれれ。

何が起きたのだろうか。なぜくだんの調査は自衛のための銃の使用を過剰に見積もったのだろうか。これは、マンモグラフィーが乳ガンにかかっていることを過剰に見積もったことと同じ理由による。偽陽性の可能性が偽陰性の可能性よりずっと高いのだ。99.9% の人が自衛のために銃を使ったことがないのに、そのうち 1% がふざけてどんな質問に対しても「はい」と答え、1% がより男らしく見せるために「はい」と答え、1% が質問内容を誤解して「はい」と答えたとすれば、自衛のための銃の使用を非常に過剰に見積もることになる。

偽陰性の方はどうだろうか？　先週強盗を銃で撃ったにもかかわらず「いいえ」と答えた人によって、偽陽性と偽陰性の数がほぼ同じとなり、相互に打ち消し合う可能性はあるだろうか。いや、そんなことはない。自衛のために銃を本当に使った人がほとんどいない場合、偽陰性となる可能性はほとんどない。偽陽性の例が偽陰性の例よりずっと多いのである。

［訳注 25］ここでは「成人の 1%」と書かれているが、正しくは、「総人口の 1%」であるべきである。アメリカ合衆国国勢調査局の推計 (`http://www.census.gov/population/estimates/nation/popclockest.txt`) によれば、1992 年 7 月 1 日のアメリカの総人口、すなわち未成年者と成人を含めた人口は、およそ 2 億 5503 万人と推計されている。その 1% は 250 万になるので、ここで触れられている銃の使用回数の 250 万回に符合するわけである。

これは先に見たガン治療薬の例に非常に類似している。ここで、$p$ は誰かが自衛のために銃を使ったことがあると間違って主張する確率である。$p$ がたとえ小さくても、最後の答えは大きく誤ったものとなるのである。

$p$ を小さくするために、犯罪学者はより詳細な調査を行う。例えば、全国犯罪被害調査[訳注 26]では、研究者が詳細な対面インタビューを行う。そのインタビューでは、回答者に対して、自衛のために銃を使用したことと犯罪について詳細にたずねる。この調査ではずっと詳しい内容が分かるので、研究者は事件が自衛に関する基準に合致しているかをよりうまく判断できる。その結果はぐんと小さなものであった。つまり、自衛のための銃の使用は百万単位で起きているのではなく、毎年 6 万 5 千件程度しか起きていない。調査に回答する人が、そうした事件を隠す可能性もあるだろうが、膨大な過剰見積もりの可能性に比べれば、ぐんと可能性が低い。

## 最初に成功しなかったら、もう一度、もう一度

基準率の誤謬は $p < 0.05$ という有意性の基準から予期されるところよりも偽陽性はずっと出やすいということを示している。けれども、ほとんどの現代の研究は有意性の検定を 1 回だけ行うわけではない。現代の研究は、最も有意な効果を探し出すべく、様々な要因の効果を比較する。

例えば、ゼリービーンズがニキビを引き起こすかどうかについて、ゼリービーンズの色ごとにニキビに対する効果を検定するとしよう。

想像できるだろうが、比較を何度も行うことは偽陽性の可能性を何度も起こすことを意味する。例えば、全くニキビを引き起こさない 20 種類のゼリービーンズフレーバーに対して検定し、$p < 0.05$ の有意度で関連性を探せば、偽陽性が得られる確率は 64% になる[訳注 27][文献 58]。45 種類の材料

[訳注 26] アメリカにおける全国犯罪被害調査 (National Crime Victimization Survey; NCVS) とは、1973 年から行われている犯罪被害に関する調査であり、アメリカ合衆国国勢調査局と司法統計局によって実施されている。

[訳注 27] 20 回の検定を行った時、偽陽性が得られる確率が 64% になることは以下のようにして

に対して検定すれば、偽陽性の確率は 90% の高きに至る。

多重比較をおかすことは簡単で、20 種類の薬の候補を試すといったことほど明白なものである必要はない。12 人の患者の症状を 12 週間にわたって追跡し、どの週でもよいから有意な利益があるかを検定してみよう。さあ、これで比較は 12 回だ。危険な副作用の候補 23 種類について、副作用が発生するか確かめてみよう。ああ、罪を犯してしまった。原子力発電所への近さ、牛乳の消費量、年齢、男のいとこの数、好きなピザのトッピング、今の靴下の色、そして他の測定しやすい要因をたくさん問うような 10 ページのアンケートを送ってみよう。何かがガンを引き起こすと発見するだろう。うんざりするほど十分な数の質問をすれば、それは不可避なのだ。

1980 年代に行われた医学に関する試験について、平均的な試験は治療上の比較を 30 回していたことを示した調査結果がある。これらの医学に関する試験の半数以上においては、研究者が多くの比較をしてしまったために、偽陽性の可能性が非常に高いものとなっている。このため、統計的に有意な結果の報告に対して疑念が投じられた。研究者は、統計的に有意な効果を発見したのかもしれないが、単に偽陽性だった可能性がある [文献 58]。

多重比較の問題を解決するテクニックはある。例えば、ボンフェローニ法 (Bonferroni correction) は、試験で比較を $n$ 回行う場合は、有意差があるとする基準を $p < \frac{0.05}{n}$ にすべきだというものだ。この方法は、偽陽性の起きる確率を、$p < 0.05$ という基準のもとで 1 回だけ比較したのと同じぐらいに下げる。だが、想像できるように、このことは検定力を下げてしまう。統計的に有意であると結論づける前に、ずっと強い相関を要求するからだ。こ

---

計算できる。まず、検定を 1 回実施したときの $p = 0.05$ というのは、偽陽性になる確率が 5% であることを示している。これは逆に言うと、偽陽性にはならない確率が 95% (0.95) であることを示している。2 回検定を実施した時に、2 つの検定が独立のものであるとすれば、2 回とも偽陽性にならない確率は、0.95 の二乗 ($0.95 \times 0.95$) で求められる。同様に、20 回検定を実施した時に 20 回全てで偽陽性にならない確率は、0.95 の 20 乗、すなわち $0.95^{20} = 0.36$ と求められる。20 回全てで偽陽性にならない確率が 0.36 (=36%) であるから、逆に言えば、1 回でも偽陽性が出る確率は $1 - 0.36 = 0.64$、すなわち 64% となる

れは難しいトレードオフだ。痛ましいことにほとんどの論文はこのことを検討しようともしないのだが。

## 脳イメージングでの燻製ニシン[訳注 28]

神経科学者は日常的に膨大な数の比較を行う。神経科学者は、fMRI[訳注 29] を使った研究をしばしば行う。そうした研究では、被験者が課題を実施する前と実施した後に、被験者の脳の 3 次元イメージが撮影される。イメージは脳内の血の流れを示し、様々な課題をした時に脳のどの部分が一番活発になるかを明らかにする。

しかし、脳のどの領域が課題をしている間に活発になるかをどうやって決めるのだろうか？　単純な方法として、脳の画像をボクセル (voxel) と呼ばれる小さな立方体に分割するものがある。課題実施前の画像でのボクセルを課題実施後のボクセルと比較し、血流の差が有意であれば、脳の部位が課題に関わっているとの結論を出すことができる。問題は比較すべきボクセルが何千とあり、偽陽性が出る可能性が非常に高いことである。

例えば、ある研究では参加者の自由回答メンタライジング[訳注 30]課題 (open-ended mentalizing task) の効果が調べられた。被験者は「規定された感情に関する値をもった社会的状況にある人間の個人を描写した一連の写真」を見せられ、「写真の中の個人はどのような感情を経験しているにちがいないかを決める」ことが求められた。この実験の間、様々な感情と論理に関する脳の中枢が光った[訳注 31]を想像できるだろう。

データが分析され、脳のある領域で、課題中に活動が変化することが分

---

[訳注 28] 燻製(くんせい)ニシン (red herring) は、話をそらすことを指す。

[訳注 29] fMRI (functional Magnetic Resonance Imaging) とは、強い磁場の中にさらすことにより、脳内の血の流れを画像の形にする手法のことである。日本語にすれば、機能的磁気共鳴画像化となる。また、その手法を実施するための装置も fMRI と呼ばれる。

[訳注 30] 他者の心の中に思い浮かんでいることを想像することをメンタライジングと呼ぶ。

[訳注 31] 脳のある領域が活性化している時、fMRI で撮った脳のイメージで、その領域は光っているように見える。

かった。イメージを比較することで、メンタライジング課題の前と後とで、脳内のとある 81 立方ミリメートルのかたまりに $p = 0.001$ の違いがあることが示された。

研究に参加した人？　いつもとは違って、10 ドルが払われる大学の学部生ではない[訳注 32]。被験者は 3.8 ポンド（およそ 1.72 kg）のタイセイヨウサケ[訳注 33]で、「スキャンをした時は生きていなかった」ものである[訳注 34][文献 8]。

もちろん、ほどんどの神経科学の研究はこれよりも洗練されている。全て一緒に変化するボクセルの集まりを探す方法や何千もの統計的検定が行われても偽陽性率を制御するテクニックがある。これらの方法は神経科学の文献では今では広く行われており、先程述べたような単純な誤りをしているような論文はほとんどない。しかし、不幸なことに、ほとんど全ての論文が、独自の方法でこの問題に対処している。241 本の fMRI の研究を調べたところ、223 種類の特有の分析戦略が用いられていることが分かった。このことは、後で議論するように、統計的に有意な結果を出すのに、研究者が非常に

[訳注 32] 神経科学や心理学では、大学生に薄謝を払って実験に参加してもらうことが多い。

[訳注 33] タイセイヨウサケは、アトランティックサーモン (Atlantic Salmon) とも呼ばれ、北大西洋と北大西洋に注ぎ込む河川に生息するサケ科の魚である（下図）。

なお、上図は Wikimedia Commons で公開されているパブリックドメイン画像を利用したものである。`http://commons.wikimedia.org/wiki/File:Salmo_salar.jpg`

[訳注 34] 脳のどの部位が働いているかどうかを調べるためには、少なくとも fMRI をかけられる生物が生きている必要がある。しかし、ここではすでに生きていないわけだから、実験として全く無意味なのである。なお、この実験をした人たちは、生きていないタイセイヨウサケの脳の活動を真面目に調べたかったわけではない。むしろ、統計の手法を濫用すれば無から有を生むことありうるということを示し、統計手法をしっかりと使うように勧めるために、あえて意味のない実験をしたのである。ちなみに、このタイセイヨウサケに対して行われた研究は、2012 年のイグノーベル賞を受賞している。

柔軟に対処できるようにしてしまっている[訳注 35][文献 13]。

## 偽発見率を制御する

先に、多重比較を修正するテクニックが存在すると述べた。例えば、ボンフェローニ法では $p < \frac{0.05}{n}$ となるところを求めることで、正しい偽陽性率を得ることができるとされている（ただし、ここで $n$ は統計的検定を実行する回数を指す）。もし、20 回の比較をする研究で、実際には存在しない効果が統計的に有意だと誤って判断する可能性を確実に 5% にとどめたかったら、$p < 0.0025$ という閾値（しきいち）を用いることになる。

これには問題がある。統計的に有意な結果があると宣言するために必要な $p$ の閾値を低くすることで、検定力を大幅に下げてしまい、真の効果を偽の効果と同じぐらい発見できなくしてしまう。ボンフェローニ法よりも、洗練された方法がいくつかある。これらの方法は検定力を向上させると言う問題のある種の統計的性質について優位に立っているが、魔法の解決手段ではない。

しかも、こうした手法は基準率の誤りの苦労から解放してくれない。$p$ の閾値にまどわされて、誤って「私が間違っている可能性は 5% しかない」と主張してしまう可能性はあるのだ。こうした手法では、偽陽性の可能性が減るだけだ。科学者がより興味を持つのは、偽発見率だ。偽発見率とは、統計的に有意な結果が偽陽性である割合のことだ。この割合を制御してくれる統計的検定はないのだろうか。

長年、この質問に対する答えは単に「ない」というものだった。基準率の誤りの節で見たように、検定された仮説のうちいくつが真であるかということについて仮定をすれば、偽発見率を計算できる。しかし、何となく推測するよりは、データから情報を見つけたい。

[訳注 35] 要するに対応方法がたくさんあるので、研究者が自分にとって都合が良い手法を恣意的に選んでしまうのである可能性が出てきてしまうのである。

1995 年、ベンジャミーニとホッホベルクは、より優れた答えを提示した。彼らはどの $p$ 値を統計的に有意なものであると考えるべきかについて見分ける非常に簡単な方法を考案した。今まで数学的に詳しいことについては触れないでいたが、この手続きがいかに簡単であるのかを示すために、数学的な話を述べようと思う。

1. 統計的検定を行い、それぞの検定について $p$ 値を求めよ。$p$ 値のリストを作って昇順に並べよ。
2. 偽発見率を選んで、それを $q$ とせよ。統計的検定の数を $m$ とせよ。
3. $p \leq \frac{iq}{m}$ となるような $p$ 値のうち最大のものを見つけよ。ただし、$i$ は並び替えられたリストの中で、$p$ 値が何番目に位置するかを示すものとする。
4. その $p$ 値とそれより小さい $p$ 値を統計的に有意であると見なせ。

できた！　この手続きは全ての統計的に有意な結果のうち $q$% を超えて偽陽性になることはないということを保証する[文献 7]。

ベンジャミーニ＝ホッホベルク法 (Benjamini-Hochberg procedure) は高速かつ有用であり、一定の分野の統計学者や科学者には広く用いられている。通常、この手法は、ボンフェローニ修正やその類似した手法に比べて検定力が良くなり、しかもより直感的な結果を返す。この手法は、様々な状況に適用可能であり、ある種のデータを検定しているとき、この手法の変種がより良い検定力をもたらす。

もちろん、これは完璧なものではない。ある種の変わった状況において、ベンジャミーニ・ホッホベルクの手法は、馬鹿げた結果を導く。そして偽発見率をコントロールすることから逃れることが常に可能であることが数学的に示されている。しかし、これは始まりだ。無いよりずっとましなのだ。

# 6 有意であるかないかの違いが有意差でない場合

訳者による概要説明
この章では、有意な結果があったものとなかったものとの間に必ずしも有意差があるわけではないということが説明されている。また、信頼区間に重複がある場合と有意差がある場合が違うことについても説明されている。

「治療法 A と治療法 B を偽薬を使った治療と比べてみた。治療法 A は偽薬に比べて有意な利点が見られたが、治療法 B は統計的に有意な利点がなかった。ゆえに治療法 A は治療法 B より優れている。」

こんな話が常にある。ここに述べられているように有意であるかないかの違いを見ることは、薬物治療、外科的処置、薬や手術によらない治療、そして実験結果を比較する簡単な方法だ。これは単純明快だ。そして、これは意味があることのように思える。

しかしながら、有意であるかないかの違いは、常に有意な差を生むわけではない［文献 22］。

セイウチの食事について比較する研究を考えてみよう。セイウチの 1 つのグループには普通の食事が与えられる。これに対して、他の 2 つのグループは新しく、より栄養のある食事が与えられる。1 ヶ月後、研究者がセイウチの体重を量ったところ、栄養のある食事 A が与えられた場合は普通の食事が与えられた場合より 25 kg 体重が重くなり、栄養のある食事 B が与えられた場合は普通の食事が与えられた場合より 10 kg しか体重が重くならなかったことが分かった。

各々の食事について平均してどれぐらい体重を増やすことが期待できるかについて立証したいとする。もし宇宙に存在する全てのセイウチにこれらの

食事を与えたら、体重増加の平均はどうなるだろうか。今、手元にそんなに多くのセイウチはいないのだから、この問題に答えることは難しい。セイウチは一頭ごとにかなり違っていて、新しい食事以外の理由で体重を増やすことがありうる。（もしかしたら、オスのセイウチは水着の季節のために大きくなっているのかもしれない。）この違いを踏まえると、食事 B の効果は統計的に有意でないと算出される。10 kg の体重増加がこの食事によって引き起こされたと結論づけるには、セイウチ間の違いが大きすぎるのだ。しかし、食事 A は統計的に有意な体重増加を引き起こしており、おそらく有効であったのだろうと考えられる。

研究者は「食事 A は統計的に有意な体重増加を引き起こし、食事 B は引き起こさなかった。明らかに食事 A の方が食事 B より太らせている」と結論づけるかもしれない。他のセイウチの飼育者はこの論文を見て、食事 A の方がより効果的だから、体重不足の病気のセイウチに食事 A を与えようと決めるかもしれない。

しかし、本当にそうだろうか。必ずしもそうではない。

なぜならば、限られたデータしかないために、数値に内在的な誤差があるのだ。他の結果についてもどれがデータと矛盾しないかを計算することは可能だ。例えば、「本当」の食事 A の効果は 35 kg の体重増加あるいは 17 kg の体重増加かもしれず、セイウチの小さな標本において、そうした結果を見ることが十分にありうる。より多くのデータを集めることで、真の効果をより正確に突き止めることができよう。

統計には、この誤差を定量化するための手法がある。各々の測定の不確かさを計算した場合、両方の食事が全く同じ効果を持つということも十分にありうることが分かる。食事 B が体重を 0 kg 増加させることは全く十分にありえることなので、食事 B の効果は統計的に有意でない。しかし、食事 B が体重を 20 kg 増加させるのに、標本の中に含まれていたセイウチが異常に痩せていた可能性も十分にありうるのだ。同様に、食事 A も 20 kg の体重増加をもたらし、研究の中で異常に食いしん坊のセイウチを使ってしまったということも全く十分にありえるのだ。より多くのデータがなければはっきりさ

せることができない。

食事 A と食事 B の間で統計的に有意な違いがあると結論づけるには、データが足りていない。片方の食事が統計的に有意な結果を出し、もう片方が出さなかったとしても、両者の間に統計的に有意な違いはない。両方とも同じぐらい有効かもしれない。2 つの結果の有意性を比べるときは気をつけよう。もし 2 つの処置あるいは効果を比較したければ、両者を直接比較しよう。

普通の文献やニュースの中でこの種の誤りの例はたくさんある。例えば、神経科学の論文は、大きな割合でこの誤りを犯している［文献 48］。数年前に、男性は生物学的な兄［訳注 36］が多いほど同性愛者となりやすいということを示唆する研究があったことを読者は覚えているかもしれない［文献 9］。どうやってこの結論に至ったのだろうか？　そしてなぜ兄であって姉ではないのだろうか？

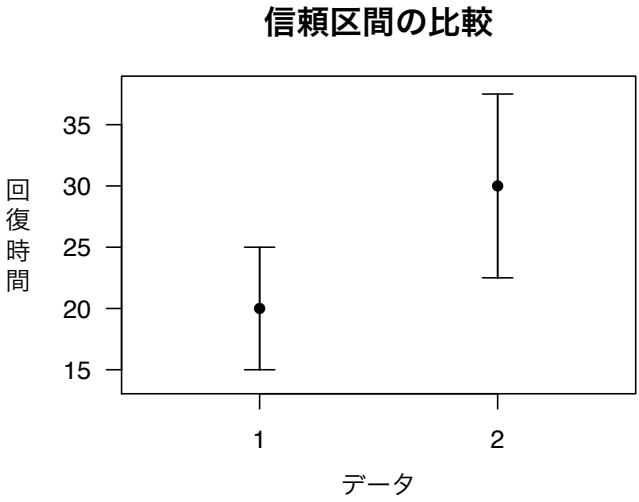
この論文の著者は、様々な要因とその同性愛への影響について分析を行うことで、結論について説明している。兄の数だけが統計的に有意な影響を示し、姉の数や非生物学的な兄の数は統計的に有意な影響を示さなかった。

しかし、今まで見てきたように、このことは兄の影響と姉の影響との間に統計的に有意な差が存在することを保証するものではない。実際、データを詳しく見てみると、兄の影響と姉の影響との間に統計的に有意な差は存在しないように見える。残念なことに、これを直接計算するに足るデータは、論文として公刊されていない［文献 22］。

## 有意差が見逃される時

この問題は他のところでも生じうる。科学者は以下のような図を用いて単に目で見るだけで有意差があるかを判断することを日常的に実施している。

［訳注 36］「生物学的な兄」とは要するに産みの親を同じにする兄のことである。




図中に描かれた 2 つの点が、各々 10 人の患者からなる 2 つの異なったグループでの、ある病気からの回復時間の推定を示しているものとしよう。図中のエラーバー[訳注 37]は 3 つの異なることを表しうる。

1. 測定の標準偏差：各々の観察点がどれだけ平均から離れているかを計算し、その差を二乗し、その結果を平均して平方根を取る。これが標準偏差で、測定がどれだけ平均から散らばっているかを測るものだ。
2. 何らかの推定量の標準誤差：例えば、エラーバーは平均の標準誤差を示すかもしれない。それぞれ別々の患者を含む標本がいくつもあり、それぞれの標本がちょうど $n$ 人の被験者を含むものとしよう。これらの標本を測定した場合、測定された平均回復時間の 68% が、「真」の平均回復時間の標準誤差 1 個分の幅に収まるということが推測できる。(平均を推定する場合、標準誤差は、測定の標準偏差を測定の数

[訳注 37] 図中で、各々の点を貫くように上下に伸びている線がエラーバーである。

の平方根で割ったものになる。だから、データを多く集めれば集めるほど、推定はより良くなる。しかし、推定が良くなるスピードはあまり早くない[訳注 38]。）例えば最小二乗回帰のように、多くの統計的手法が、結果に対する標準誤差の推定を与える。

3. 何らかの推定量の信頼区間：95% 信頼区間とは、100 個のランダムな標本の中から 95 のランダムな標本について真の値を含むように数学的に構成されたものである。だから、およそ各々の方向に大体標準誤差 2 つ分だけ広がっていることになる。（より複雑な統計モデルにおいては、これは正確でないかもしれないが。）

これら 3 種類は全て異なっている。標準偏差は自分のデータに対する単純な測定である。標準誤差は、平均や最も当てはまりの良い直線の傾きといった統計量について、患者の標本をいくつも取ったとしたら、どれぐらいの幅が見られるだろうかを教えてくれるものである。信頼区間は標準誤差に類似しているが、95% 信頼区間の 95% が「真」の値を含んでいるに違いないということを保証してくれる点で標準誤差と異なっている。

上述の図の例では、重なり合う 2 つの 95% 信頼区間がある。多くの科学者は、これを見て 2 つのグループの間に統計的に有意な差はないと結論づけるだろう。結局、グループ 1 とグループ 2 は違わないのかもしれない。平均回復時間は両方のグループで 25 かもしれない。そして、今回はグループ 1 が幸運であったために、違いが表れただけかもしれない。しかし、このことは差が統計的に有意であることを示すだろうか？　$p$ 値はどうなるだろうか？

この場合、$p < 0.05$ となる。たとえ、信頼区間が重なり合っていたとして

[訳注 38] ここで、標準誤差の値が小さいほど、推定のぶれの幅が小さいということになる。標準誤差はデータの数の平方根の逆数に比例するので、データを 2 倍に増やしても、標準誤差の範囲は 70.7% にしか狭まらない。データの数が 3 倍になっても 57.7%、データの数が 4 倍で 50.0% といったところだ。その意味で推定が良くなるスピードは「早くない」のである。

も、グループの間には統計的に有意な差があるのだ[原注 2]。

残念なことに、多くの科学者は仮説検定を省き、信頼区間が重なっているかを確認するために、プロットをちらっと見るだけで済ましてしまう。これは実際には非常に保守的な検定である[訳注 39]。信頼区間が重ならないように要求することは、状況によっては $p < 0.01$ を要求すること似たようなことになる[文献 54]。2 つの測定の間に有意な差があるにもかかわらず、ないと主張するのは簡単なのだ。

逆に言えば、測定を標準誤差や標準偏差で比較することも誤解を招きやすい。標準誤差の幅は信頼区間の幅より狭いからだ。2 つの観測結果で、標準誤差が重なり合わないが、両者の違いは統計的に有意でないことはありうるのだ。

心理学者・神経科学者・医学研究者に対する調査で、大多数の人がこの単純な過ちを犯していること、そして多くの科学者が標準誤差・標準偏差・信頼区間を混同していることが分かっている[文献 6]。気候科学の論文に対する他の調査では、2 つのグループをエラーバーで比較している論文の大部分がこの過ちを犯していることを発見している[文献 40]。『誤差分析入門』といった実験科学者のための入門教科書でさえ、学生に対して目で見て判断するように教えており、正式な仮説検定を全くほとんど触れないでいる。

もちろん、目で見て比較できる信頼区間を生成する正式な統計手続きは存在する。その手続きは自動的に多重比較を修正してもくれる。例えば、ガブリエル比較区間 (Gabriel comparison interval) は目で見て簡単に解釈される[文献 19]。

---

[原注 2] これは、グループ 1 の標準誤差は 2.5、グループ 2 の標準誤差は 3.5 であるということに基づいて、対応なしの $t$ 検定で計算されたものである。

[訳注 39] 本文で述べられているように、通常、仮説検定と信頼区間だと、信頼区間の方がグループ間の差があると認定するハードルが高い。だが、これを逆から見れば、$p$ が 0.05 より大きいか小さいかで判断する場合、仮説検定はグループ間に差があると認定する基準が緩すぎるという話になる。このため、$p$ が 0.05 より大きいか小さいかという緩い基準で仮説検定は用いるべきではないと主張している人もいる[文献 36]。

信頼区間が重なっていることは、2 つの値に有意差がないことを意味しない。同様に、標準誤差のバーが重なっていないことは、2 つの値に有意差があることを意味しない。代わりに適切な仮説検定を用いることが常に最良の手段である。あなたの眼球はちゃんと定義された統計的手続きではないのだ。

### ❖ 訳者コラム：$p$ 値に関する様々な誤解 ❖

$p$ 値は誤って解釈されることが多い。グッドマンは $p$ 値に関して以下の 12 個の誤解を挙げている［文献 25］。

1. $p = 0.05$ ならば、帰無仮説が真である確率は 5% しかない。
2. $p \geq 0.05$ のような有意でない結果は、グループ間に差がないことを意味する。
3. 統計的に有意な発見は客観的に重要である。
4. $p$ 値が 0.05 より大きい研究と小さい研究は矛盾する。
5. $p$ 値が同じ研究は帰無仮説に対して同等の証拠を提供する。
6. $p = 0.05$ は、帰無仮説のもとで 5% しか起こりえないデータを観察したことを意味する。
7. $p = 0.05$ と $p \leq 0.05$ は同じことである。
8. $p$ 値は不等式の形で書かれるべきものである（例えば、$p = 0.015$ のときは $p \leq 0.02$ とする）。
9. $p = 0.05$ は、帰無仮説を棄却したとしたら、第一種の誤りの確率が 5% しかないことを示す。
10. 有意水準 $p = 0.05$ の下で、第一種の誤りの確率は 5% になる。
11. ある方向を向いた結果やその方向の結果がありえない差異を気に留めないのであれば、片側の $p$ 値を用いるべきである。
12. 科学に関する結論や処置の方針は $p$ 値が有意であるかどうかに基づくべきである。

# 7 停止規則と平均への回帰

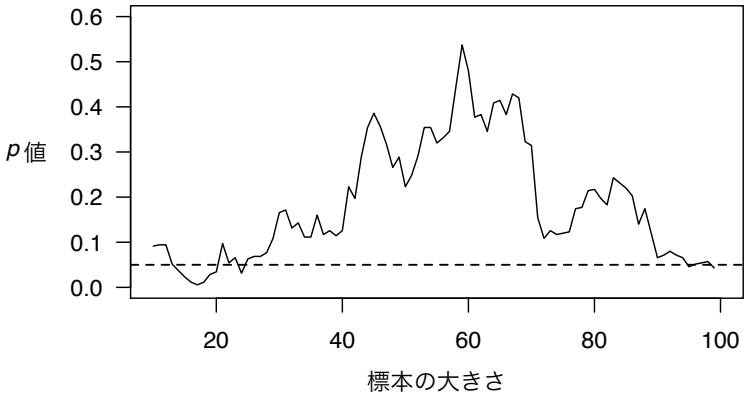
訳者による概要説明
この章では、研究を途中で中止することによって発生する問題、結果が実際よりも誇張されてしまう問題について扱われている。この2つの問題は別々のものでなく、相互に関係していることに注意が必要である。さらに、この章の最後の方では、規模が小さいグループほど平均からずれやすく、規模が大きくなるに連れて平均に近づいていく（＝回帰する）ことが紹介されている。

医学に関する試験は多額の費用を要する。たくさんの患者に対し、実験的な薬物治療を実施し、何ヶ月もの間、その症状を追跡することは、相当な量の資源を消費する。このため、多くの製薬会社が「停止規則」(stopping rule) を発達させてきた。これは実験的な薬が実質的な効果を持つと明らかになった場合、調査者が研究を早めに終えることを許す規則である。例えば、試験は半分しか終わっていないものの、新しい薬について、症状に統計的に有意な差がすでに存在するならば、研究者は結論をより強固なものにするためにより多くのデータを集めるのではなく、研究を終わらせてもよい。

しかし、下手に行われれば、このことは偽陽性を多数もたらしてしまう可能性がある。

例えば、2つのグループの患者を比較するとしよう。ただし、一方のグループには薬が投与され、もう一方のグループには偽薬が投与されるものとする。薬が働いているかを調べるために、あるタンパク質について血流中での濃度を測定する。ただ、ここで、この薬は全く違いをもたらさない。もちろん、人によって濃度は少し異なるだろうが、2つのグループで、患者のタンパク質の濃度の平均は同様である。

各々のグループについて 10 人の患者から始め、徐々に、より多くの患者からより多くのデータを集めるものとする。進める際に、2 つのグループを比較するために $t$ 検定を行い、平均のタンパク質濃度の間に統計的な有意差があるかを見る。このシミュレーションのような結果が得られることになるだろう。




この図は、より多くのデータを集めたときのグループ間の差異についての $p$ 値を示している。水平線は $p = 0.05$ の有意水準を示している[訳注 40]。最初は、有意差がないように見える。そして、より多くのデータを集めて、有意差があると結論づける。もし中止したとしたら、誤ったことを信じていただろう。グループ間の有意差が本当は存在しないにもかかわらず、有意差があると信じていただろう。さらに多くのデータを集めたら、間違っていたということに気づく。結局、ちょっとした幸運が偽陽性へと導いてしまっていたのだ。

グループの間で本当の違いがないのだから、$p$ 値が一時的に小さくなることが起きるわけがないと思うかもしれない。つまり、より多くのデータを取

[訳注 40] この水平線より下に来ていれば、$p < 0.05$ となり、有意差があると結論づけてしまう。

ることで、結論をよりダメなものにしてしまうわけがないというわけだ。そうだよね？　再度試験を行った場合、最初からグループ間に有意差がなく、より多くのデータを集めてもそのまま有意差がないままでいるというのはありえる。また、巨大な差が存在する状態で始まり、即座に差がない状態に戻るというのもありえる。しかし、もし充分長く待ちつつ、データポイントが1つ加わるごとに検定するならば、たとえ本当は差が全くなくても、任意の値の統計的有意水準を下回ることがあるだろう。通常、無限の標本を集めることはできないので、現実にはこうしたことが常に起きるわけではないが、そうだとしても、うまく実施されない停止規則は偽陽性率を大きく上昇させる［文献 57］。

現代の臨床試験では、統計に関する実験計画をあらかじめ登録することがしばしば求められる。そして、一般的には、観察1つが終わったらそのつど検定するのではなく、証拠を検定するための少数の評価点を先に選んでおく。このことは偽陽性率を少ししか引き上げず、しかもここでの偽陽性率は、必要な有意水準を注意深く選び、より進んだ統計的技法を用いることで調整することができる［文献 61］。しかし、実験計画が登録されず、研究者が適切だと感じる手法を何でも使える自由がある分野では、偽陽性の悪魔が潜んでいるかもしれない［訳注 41］。

## 真実の誇張

医学に関する試験は、薬の間の中程度の差異を検出するための検定力が十分でない傾向にもある。だから、効果を見つけたらすぐにやめたくとも、効果を検出するのに十分な検定力がないのだ。

［訳注 41］ベイズ統計の手法に逐次確率比検定というものがある。この手法は、データを増やすたびに検定を行う手法であるが、普通の検定と違って、(1) 有意差があると結論づける、(2) 有意差がないと結論づけるという2つの選択肢の他に、(3) 判断を留保してもっとデータを集めるという選択肢が用意されている。このため、本文で述べられたような問題が起こりにくくなっている。

ある薬が、偽薬に比べて 20% 症状を減らすとしよう。しかし、検定するために実施している試験は、この差異を検出するための十分な検定力がないものとしよう。小規模の試験では、幅広い結果が出る傾向があることが知られている。普通より短いかぜをひいている 10 人の幸運な患者を得ることは簡単だが、みんな普通より短いかぜをひいている 1 万人を得ることは、ずっと難しいのだ。

この試験を何回も実施することを想像してみよう。時には不幸な患者を得て、薬から得られる統計的に有意な改善に気づかないこともある。時には患者がちょうど平均的で、実験群では症状が 20% 減少する――だが、これを統計的に有意な増大と呼ぶには十分なデータがないので、これを無視する。時には患者が幸運で、症状が 20% よりずっと多く減少する。そして、試験を止めて、「ね！　うまくいっただろ！」と言うことになる。

薬が有効であると正しく結論づけることができたが、その効果の大きさを誇張してしまった。この薬が実際よりもずっと有効であると誤って信じてしまったのである。

こうした現象は、薬物試験、疫学研究、遺伝子関連解析（「遺伝子 A が状況 B を引き起こす」）、心理学研究、そして医学文献の中で最も良く引用されている論文で発生している［文献 32,34］。（遺伝子関連解析のように）多くの独立した研究者によって試験が短時間で行われる分野では、最も初期に公刊される結果はしばしばきわめて矛盾したものとなる。小規模試験と統計的有意性を求めることとによって、最も極端な結果しか公刊されないことになるからだ［文献 35］。

おまけに、真実の誇張は早期停止規則によって引き起こされうる。もし臨床試験でほとんどの薬が、試験を早期に止めることを保証できるほど有効でない場合、早期に停止される試験の多くは、幸運な患者と大したことのない薬の結果ということになるだろう。そして、試験を停止することは、差を判断するのに必要な追加のデータを自身から奪っていることになる。早期に停止した試験と、同じ課題に取り組んで早期に停止しなかった他の試験とを比べた報告がある。この報告によると、ほとんどの場合、早期に停止した試験

は、試験対象となった治療の効果を誇張しており、その誇張の度合いは平均して 29% であった［文献 3］。

もちろん、研究対象となっているいかなる薬についても「真実」は知りようがない。だから、早期に停止した研究が、幸運によるものなのか特に良い薬によるものなのかを判断することはできない。多くの研究は、元々意図していた標本の大きさ (sample size) や研究を終わらせることを正当化するために用いられた停止規則さえ公表しようとしない［文献 47］。試験が早期に止められていることは、その結果が偏っていることを自動的に示す証拠ではない。しかし、偏っていることをほのめかすものである。

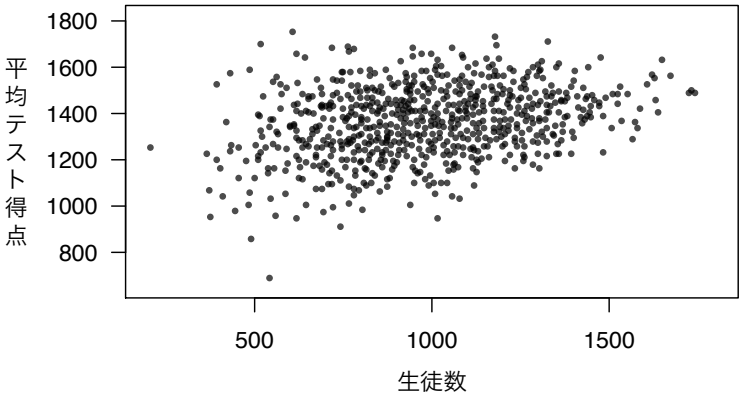
## 小さな極端なもの

あなたが公立学校の改革を担当しているとしよう。最も良い教授方法についての調査の一部として、あなたは、学校の大きさが標準化されたテストの得点に与える影響を見ている。小さい学校の方が大きい学校よりうまくいっているのだろうか？　あなたは小さな学校をたくさん建てるべきだろうか？　それとも大きな学校を少しだけ建てるべきだろうか？

この問題に答えるため、成績上位の学校のリストを集めた。平均的な学校には千人程度の生徒がいるが、成績上位の 5 校あるいは 10 校はほとんど全てそれより生徒の数が少なかった。このことは、小さい学校が最も良くやっているように見える。もしかしたら、教師が生徒を知ることができて、個別に助けることができる個人的な雰囲気によるものかもしれない。

そして、あなたは、成績下位の学校は何千人もの生徒と働きすぎの教師がいる都会の大きな学校だろうと予測しつつ、成績下位の学校を見てみた。なんと！　成績下位の学校もみんな小さな学校だった。

何が起きているのだろうか？　テスト得点と学校規模の関係を示した図を見てみよう。




学校の中でも小規模なものほど、テストの平均得点が大きくばらついている[訳注 42]。これは、こうした学校に生徒が少ないために散らばっているのに他ならない。生徒が少なければ、教師の「真」の能力をはっきりさせるデータ点[訳注 43]が少ないということになる。だから、平均得点は大きくばらつくのだ[訳注 44]。学校が多くなるほど、テスト得点のばらつきは少なくなる。そして、実は平均してテスト得点は増加しているのだ。

[訳注 42] 図の左側の方に表されているのが小規模の学校である。小規模の学校のテスト得点は上から下までばらついている。これに対して、図の右側の方に表されている大規模の学校は、テスト得点がそれほど上下にばらついていない。

[訳注 43] 個々の生徒の得点は、それぞれが教師の教える能力を反映したデータ点として捉えることができる。

[訳注 44] これは今まで触れてきた薬の効果の試験と全く同じ話である。薬について調べる際、わずかな数の患者に対してしかデータを取らなかった場合のことを思い出してみよう。例えば、10 人しか患者がいなければ、薬が効果を及ぼさない場合でも、全員がたまたま軽い症状だったという可能性は低くない。逆に全員がたまたま重い症状だったという可能性も低くない。つまり、人数が少ないと、極端な結果が出やすい。もし 1 万人の患者がいれば全員がたまたま症状が軽かったり重かったりすることは少ないだろう。つまり、人数が多いと、極端な結果が出にくくなる。学校の生徒の人数とテスト得点の話も、この薬の試験の話と同じである。人数が少ない方が極端な結果が出やすく、多い方が出にくいのである。

この例はシミュレーションで作られたデータを用いている。しかし、この例は、ペンシルベニアの公立学校の実際の（そして驚くべき）観察結果に基づいて作られたものである［文献 64］。

他の例を挙げよう。アメリカでは、腎臓ガン罹患率の最も低い部類の郡は、中西部・南部・西部の田舎の郡である傾向がある。どうしてこうなるのだろうか？　様々な説明を考えることができるだろう。田舎の人は、運動量が多く、汚染の少ない空気を吸い、そしてもしかしたらストレスの少ない生活をしているのかもしれない。こうした要因がガン罹患率を下げているかもしれない。

これに対して、腎臓ガン罹患率が最も高い部類の郡は、中西部・南部・西部の田舎の郡である傾向がある。

もちろん、田舎の郡が人口がとても少ないことに問題のカギがある。10 人の住民しかいない郡［訳注 45］で、腎臓ガン患者が 1 人いれば、その郡が国内でもっとも腎臓ガンが高い郡となってしまう。つまり、小さな郡は、単に住民が非常に少ないために、腎臓ガン罹患率が非常にばらつくのである［文献 21］。

［訳注 45］アメリカの郡 (county) は大きいものから小さいものまで様々あるが、さすがに 10 人しか住民のいない小さな郡はない。ただし、数百人しか住人がいない小さな郡ならば多数存在する。なお、アメリカにはおよそ 3,000 の郡があり、1 郡の住民数は平均 10 万人程度である。

# 8 研究者の自由：好ましい雰囲気？

訳者による概要説明
この章では、実際に統計的な分析を行う際、どう分析するかについて決定すべきことが多数あるということが述べられている。ただし、あまり気ままに分析を行うことが許されれば、研究者にとって都合の良い結果が出てくるだろうという警告も述べられている。

統計は退屈で単調なものだという広く知られた誤解が存在する。たくさんのデータを集めて、数を Excel とか SPSS[訳注 46] とか R[訳注 47] とかにつめこんで、そしてソフトがカラフルな図を出力するまで棒でたたく。おしまい！　統計家のやるべきことは、結果を読み上げるだけだ。

だけれども、どのコマンドを使うかについては選ばなくてはならない。2人の研究者が同じ問題に答えるために、全く違った統計分析をすることもありうる。決めなくてはならないことがたくさんあるのだ。

1. どんな要因を調節するかを決めなくてはならない。例えば、医学に関する試験ならば、患者の年齢、性別、体重、BMI、以前の病歴、喫煙の有無、薬の使用の有無、あるいは研究の前に行われた医療検査の結果などを統制[訳注 48]することになるかもしれない。これらの要因のうち、どれが重要で、どれが無視できるものかということも決めなく

[訳注 46] SPSS は統計解析を行うソフトの 1 つであり、今は IBM が販売している。
[訳注 47] R は、統計解析向けのプログラミング言語の 1 つである。
[訳注 48] ここでの「統制」(control) とは、調査する対象に偏りがないように調節することを指す。例えば、薬の効果を調べる際に、男女を問わずに使える薬であると仮定しているならば、被験者が男ばかりになってしまうのは、統制が取れていないということになる。ちゃんと統制する方法として、例えば、男女を半々にして性別の偏りを防ぐことが考えられる。

てはならない。

2. どんな事例を除外するかを決めなくてはならない。食事のプランを試しているときに、コントロールできない下痢で倒れてしまった被験者がいたら、結果が正常なものにはならないから、その被験者を除外したいと考えるかもしれない。
3. 外れ値 (outlier) にどう対処するかを決めなくてはならない。理由が分かるものにせよ分からないものにせよ、普通のものから外れてしまっている結果というものは常にあって、そうしたものを除外したり、特別に分析したりしたいかもしれない。どんな事例を外れ値と見なして、そしてどう対処すべきなのかを決めなくてはならない。
4. グループをどう定義するかを決めなくてはならない。例えば、患者を「肥満」・「正常」・「痩せ」というグループに分けたい時、どこに境界を設定すべきか決めなくてはならないし、BMI が「肥満」の範囲に入ってしまっているむきむきのボディビルダーについてはどうすれば良いか決めなくてはならない。
5. 欠損データ (missing data) についてどうすべきかを決めなくてはならない。新しい薬で、ガンの寛解[訳注 49]率を試験することがあるかもしれない。5 年に及ぶ調査を実施するとしても、6 年後あるいは 8 年後に腫瘍が再び出現する患者がいるかもしれない。データの中にはこうした病気の再発が含まれない。薬の有効性について測定する際に、このことについてどう説明すべきかを決めなくてはならない。
6. データをどれだけ集めるべきかを決めなくてはならない。自信が持てる結果が出たらデータ収集をやめるべきか、全てのデータが集まるまで計画したどおりのデータ収集を続けるべきかを決めなくてはならない。
7. 結果をどう測定するかを決めなくてはならない。薬は、患者の主観に

[訳注 49] 病気の症状がほぼ消えることを寛解と呼ぶ。症状が問題ない程度になっているだけで、完全に治癒されたとは言えない状況についても寛解に含まれる。

基づく調査でも評価できるだろうし、医学検査の結果でも、ある症状の罹患率でも、病気の継続期間などの基準でも評価できるだろう。

結果を得るために、どの手続きが最も適切かを見る探求・分析が何時間もかかるだろう。論文では、実施した統計分析についての説明を通常行う。しかし、なぜ研究者がある方法を選んで他の方法を選ばなかったかということについてはいつも説明するわけではないし、他の方法を選択した場合どんな結果が得られただろうかということについても説明するわけではない。研究者は自身が適切だと感じるものを何でも選ぶ自由がある。そして、研究者は正しい選択をするかもしれない。だが、データに対して異なった分析をした場合、どうなるだろうか。

シミュレーションによれば、単純に、異なった変数を調整したり、異なった事例のセットを排除したり、外れ値の扱いを変えたりすることで、2 倍の違いがある効果量[訳注 50]を得ることができる［文献 34］。効果量というのは、薬が引き起こす違いがどれくらいかを教えてくれる、例のきわめて重要な数字のことだ。だから、どうやらやりたいように分析する自由があれば、結果を大いにコントロールすることができるようなのだ[訳注 51]。

統計の自由による最も気がかりな影響は、研究者が自分にとって一番都合の良い統計分析を選んでしまうことだ。何かが出てくるまでデータをいじくりまわすことで、統計的に有意な結果を恣意的に生み出すのだ。与えられたデータセットに対してうまくやれる手法が見つかるまで異なった統計分析手法を研究者に試させつづけるだけで、偽陽性率は 50% に跳ね上がりうるということが、シミュレーションによって示唆されている［文献 57］。

---

［訳注 50］効果量 (effect size) とは、効果の大きさを表した量のことである。先に見たように、*p* 値は効果の大きさを表した量ではない。それにもかかわらず、効果量でなく、*p* 値を効果の大きさを表すのに使ってしまう統計の誤用はしばしば見られる。統計解析を行う人は、これら 2 つの概念を混同しないように注意する必要がある。

［訳注 51］一般的に効果量が大きければ大きいほど、有意であるという結果が出やすくなる。よって、効果量が大きくなる方向に持っていくようにすれば、有意な結果を簡単に出すことができる。

医学の研究者はこういったことを防ぐ手法を工夫してきた。データがどのように集められてどのように分析されるのかについて説明するために、臨床試験のプロトコルの草稿を出すことが研究者にしばしば求められる。研究者がデータを見る前に草稿が出されたプロトコルだから、自分にとって一番都合の良い分析をこねくりだせるわけがない。残念なことに、多くの研究ではプロトコルを逸脱して、異なった分析をし、研究者のバイアスが入り込みうるようになってしまっている［文献 14,15］。他の多くの科学の分野では、プロトコル公表が要件として課されることは全くない。

統計の手法が増えることは、様々な道具立てをもたらしてくれる。しかし、統計の手法は鈍器のように用いられているようにも見える。データが白状するまで、データを単にたたいている人がいるに違いないのだ。

# 9 誰もが間違える

訳者による概要説明
この章では、現在の科学研究において、統計の誤りが多いことについて説明した上で、こうした誤りに対抗するにはデータの共有が重要だと訴えている。

今までの議論では、科学者は計算するための適切な数字を選ぶことを間違えるだけで、統計に関する計算を完全に正しくできるものだと考えてきた。科学者は統計的検定の結果を誤って使ったり、関連する計算に失敗するかもしれないけれども、少なくとも $p$ 値は計算できる。良いよね？

たぶんそうじゃない。

医学と心理学の実験で報告された統計的に有意な結果に対して調査を実施したところ、多くの $p$ 値が間違っていることが示された。また、統計的に有意でない結果についてちゃんと計算したところ、本当は有意であるものが存在することが示された［文献 2,27］。他の報告では、誤って分類されたデータ、間違って重複してしまったデータ、おかしなデータセットをまるごと入れること、そしてその他の混乱の事例が示されている。こうした事例は、間違いについて簡単に気づくように十分な詳細を記述しなかった論文では全て隠されている［文献 1,26］。

殺菌するものの中では日光が最も良い［訳注 52］。つまり、みんなに注目さ

［訳注 52］原文は “Sunshine is the best disinfectant”（日光は最も良く殺菌するものである）となっている。これは、アメリカの法律家ルイス・ブランダイスの “Publicity is justly commended as a remedy for social and industrial diseases. Sunlight is said to be the best of disinfectants; electric light the most efficient policeman.”（公（おおやけ）に注目されることは、社会・産業の病（やまい）の治療法として正当に推奨される。日光は殺菌するものの

れることが問題を解決する良い手段なのだ。そして、多くの科学者は、実験データがインターネットを通じて手に入れられるようにすることを求めている。いくつかの分野では、こうしたことがありふれたことになっている。遺伝子配列データベース、タンパク質構造データバンク、天体観測データベース、地球観察コレクションといった多くの科学者の貢献が含まれているデータが存在している。しかし、他の多くの分野では、データを共有できないでいる。量子力学のデータではテラバイト単位の情報を含むといったように、実用的な理由から共有できないこともある。医学実験のようにプライバシーの問題があるから共有できないこともある。また、資金や技術的サポートがないために共有できないこともあるし、あるいは単にデータとそこから得られる結果の全てを独占的なコントロールのもとに置きたいと思っているためにデータを共有できないこともある。そして、たとえデータが全て手に入ったとしても、誰が誤りを見つけるために分析するだろうか？

同様に、ある種の分野の科学者は、うまくできた技術的ツールを使って統計分析の内容を手に入れられるようにしている。例えば、Sweave というツールでは、科学・数学の出版で標準的となっている LaTeX というもので書かれた論文の中に、行われた統計の結果を人気のある R 言語を使って簡単に埋め込むことができる。結果は普通の科学論文と同じように見えるが、その論文を読んでその手法に興味を持った他の科学者がソースコードをダウンロードすることができる。そのソースコードには全ての数値がどう計算されたかが書いてある。しかし、科学者はこうした機会を利用するだろうか？コードの誤字をチェックしても、誰も科学における栄誉は得られないのだ。

他の解決方法としては繰り返し (replication) があるだろう。科学者が他の科学者の実験を注意深く再現して結果を検証するのなら、誤った結果を引き起こす誤字の可能性を除外するよりずっと楽だ。繰り返しはめったに起きない偽陽性の結果も除外する。多くの科学者は実験の繰り返しは科学の真髄

---

中で最も良いと言われている。電灯は最も効率的な警察官だ。）という名言を引いたものである。

であると考えている。新しい考えは、それが独立に試験され、世界中で再試験が行われ、筋が通っていると分かるまで認められないのだ。

このことは完全に正しいわけではない。科学者はしばしば先行研究を正しいものだと考える。だが時には過去の研究成果について系統的に再試験をしようと決めることがある。例えば、ある新しいプロジェクトは、主要な心理学誌に載った論文の再現をすることを目的としている。そこでは、論文のどれだけが今なお有効なのかをはっきりさせ、論文のどのような特性が再試験に耐えることができるかを予測できるかをはっきりさせようとしている[原注 3]。他の事例として、アムジェン[訳注 53]のガンの研究者たちが 53 のガン研究における画期的な前臨床研究について再試験を行ったことがある。(「前臨床」という言葉は、研究が新しくて未証明の考えについて試験しているために、人間の患者には関わらなかった研究[訳注 54]であるということを意味している。) 原論文の著者と協力したにもかかわらず、アムジェンの研究者は、再試験をした研究のうち 6 つでしか結果を再現することができなかった[文献 5]。バイエル[訳注 55]の研究者は、公刊された論文の中で見つかった新しい薬として使える可能性がある薬の試験をした際に、同様の困難を報告している[文献 53]。

これはやっかいだ。この傾向はより理論的でない医学研究にも当てはまるだろうか。どうもそうらしい。医学で最も良く引用されている研究記事のうち、4 つに 1 つが記事出版後に再試験が行われていないし、3 つに 1 つが後の研究で誇張されたものか誤っているものであると分かっている[文献 32]。これはアムジェンの結果ほど極端ではないが、重要な研究の中にどんな誤り

[原注 3] 再現性プロジェクト (The Reproducibility Project)：
`http://openscienceframework.org/reproducibility/`

[訳注 53] アムジェン (Amgen) はアメリカのバイオテクノロジー企業で、医薬品の開発・製造を業務としている。

[訳注 54] 新しいものをいきなり人間の患者に実施するのは危険である。このため、先に動物実験を実施して安全性を確認するということなどが行われる。

[訳注 55] バイエル (Bayer) はドイツの製薬会社である。

が気づかれないまま潜んでいるのだろうかという疑いを持たせるだろう。繰り返しは我々が期待しているほど広く行われてはいない。そして、結果はいつも歓迎すべきものであるとは限らないのだ。

# 10 データを隠すこと

訳者による概要説明
この章では、科学者がデータを共有したがらないために発生する問題点について述べている。

> 十分な数の目玉があれば、全てのバグは大したものではない。
>
> ——エリック・スティーブン・レイモンド[訳注 56]

科学者が犯しがちな誤りについて述べてきた。そして、外部からの少々の監視の目がこうした誤りを発見するためにどれほど最高の手段かについて述べてきた。査読はこうした監視の目を多少はもたらす。しかし、査読者にはデータを広範囲にわたって再分析したり、コードの誤字を見る時間はない。査読者は方法論が筋が通っているかどうかだけをチェックする。時には明らかな誤りを発見することもあるが、微妙な問題については通常見逃される[文献 56]。

多くの学術誌や専門の学会が研究者にデータを他の科学者に対して提供できるように求めているのはこのためだ。完全なデータセットは通例学術誌のページに印刷するには大きすぎる。だから、著者は結果を報告した上で、も

[訳注 56] エリック・スティーブン・レイモンド (Eric Steven Raymond) はアメリカの有名なプログラマである。『ハッカーズ大辞典』の編者としても知られている。レイモンドは、「十分な数の目玉があれば、全てのバグは大したものではない。」(“Given enough eyeballs, all bugs are shallow.”) という言葉によって、ソフトウェア開発において多数の人の目にさらされればソフトウェアの不具合は修正されるということを簡潔に言い表している。レイモンドの言葉はソフトウェア開発に関する話であるが、この考え方は科学研究にも応用できるというのがこの章の主眼である。

しコピーを求められれば完全なデータを他の科学者に送る。もしかしたら他の科学者が誤りや元の研究をした科学者が見落としたパターンに気づくかもしれない。

理論上はそれでうまくいくのかもしれない。2005 年、アムステルダム大学のイェルテ・ヴィヒェルツは同僚とともに、アメリカ心理学会[訳注 57]のいくつかの重要な学術誌に出ている最近の記事の全てを分析しようと決めた。それらの記事で使われている統計手法を知るためである。アメリカ心理学会は、論文の著者に対して、著者の主張を検証しようとする他の心理学者にデータを共有することを求めている。これが同学会を選んだ理由の一つである。

ヴィヒェルツたちがデータを求めた 249 個の研究のうち、6 ヶ月以内にデータを受け取れたのは、64 個だけだった。全体の 4 分の 3 近くの研究で、著者がデータを全く送ってこなかったのである［文献 66］。

もちろん科学者は忙しい人種だから、データセットをまとめて、各々の変数が何を意味していてどう測られたかといったことを記述した文書を作る時間がなかっただけなのかもしれない。

ヴィヒェルツとその同僚は、これを調べることを決意した。首尾一貫しない統計の結果や様々な統計的検定の誤用、一般的な誤字といった論文を読むことで見つけることができるありふれた誤りを探すために、全ての研究を調査した。少なくとも半分の論文で誤りが 1 つはあった。たいていは小さな誤りであったが、15% は誤りのせいで統計的に有意になっているだけの「有意」な結果を少なくとも 1 つは報告していた。

次に、こうした誤りとデータを共有したがらないこととの関係について探索したところ、両者の間に明らかな関係があった。データを共有することを

［訳注 57］アメリカ心理学会 (American Psychological Association; APA) は、その名の通り、アメリカの心理学者が集まってできた学会である。その規模は非常に巨大であり、13 万の会員を擁する。同学会は最も代表的な *American Psychologist* のほか、感情研究を扱う *Emotion* や教育心理学を扱う *Journal of Educational Psychology* など、心理学の様々な分野について学術誌を出している。

拒絶した著者は、論文の中で誤りを犯しがちで、統計的な証拠が弱くなりがちな傾向があった［文献 65］。ほとんどの著者がデータを共有することを拒否したから、ヴィヒェルツは統計的な誤りを深く掘り下げることができなかったが、より多くの誤りが潜んでいるかもしれない。

決してこれは作者が誤りを発見されるのを恐れてデータを隠していたり、誤りについて知っていたということの確実な証拠ではない。相関関係は因果関係を含意しない。しかし、相関関係は、示唆的に眉を揺らして、こっそりジェスチャーをしつつ、声を出さずに口だけを動かして「あそこを見ろ」と言うのだ。［原注 4］

## 詳細は省略しておけ

あらさがしをする統計学者が論文の欠陥を指摘してげんなりさせるって？分かりやすい解決方法が 1 つある。あまり詳細を公表しないことだ！　データをどう評価したかを言わなければ、統計学者は誤りを見つけることができないのだ。

悪意ある科学者がこうしたことを意図的に行っていると本気で言うつもりはない。もしかしたら、そういう科学者もいるかもしれないが。より頻繁に起こるのは、単に著者が詳細を含めることを忘れてしまったせいで、詳細が載らないことだ。あるいは、学術誌のスペースが限られているために、割愛せざるを得なかったということだ。

載せなかったのが何かを見るべく研究を評価することは可能である。医学に関する試験を主導する科学者は、試験を始める前に倫理審査委員会［訳注 58］に詳細な研究計画を提示することが求められる。そして、ある研

［原注 4］これは、恥知らずにも `http://xkcd.com/552/` の代替テキストから盗用したジョークである。

［訳注 58］医学に関する試験を実施する大学や病院などでは、倫理審査委員会 (ethical review board) が設置される。倫理審査委員会は、試験を行う人から独立して、試験が倫理的に

究者グループはこうした計画を集めたものを委員会から手に入れた。計画においては、研究でどの結果を測定するのかということが具体的に述べられている。例えば、ある研究では、治療によって何か影響を受けた症状があるかを見るために、様々な症状をチェックするかもしれない。そして、くだんの研究者グループはこれらの研究の出版された結果を見つけて、これらの成果がどれだけしっかりと報告されているかを調べた。

成果のおよそ半数が、学術誌に載った論文に全く出ていなかった。これらのほとんどは、統計的に有意でない成果で、ゴミをほうきで掃いてじゅうたんの下に入れたか[原注 5][訳注 59]のように、隠されていたものである。また、結果のその他のかなりの部分は、さらなるメタ分析[訳注 60]を行うために結果を使おうとする科学者にとって十分な詳細が報告されていなかった[文献 14]。

他にも同様な問題が報告されている。医学に関する試験についてのある報告では、ほとんどの研究が停止規則や検定力の計算といった重要な方法論に関する詳細を省略していることが示されている。大きな一般的な医学誌に比べて、小さな専門的な学術誌に載っている研究の方がまずいことになっている[文献 31]。

医学誌は、CONSORT チェックリスト[訳注 61] のような結果報告の基準

[原注 5] なぜ我々はいつも「ほうきで掃いてじゅうたんの下に入れる」と言うのだろうか。それは誰のじゅうたんなのだろうか。そして、なぜほうきのかわりに掃除機を使わないのだろうか

問題ないかについて判断することが求められる。医学に関する試験を行う際には、試験中の薬による副作用など、被験者に悪影響が及ぶ可能性がある。こうした問題を防ぐために、試験を始める前に倫理審査委員会の審査と承認を経ることが必要とされる。

[訳注 59] 英語では「(悪い物事を) 隠す」という意味で、“sweep under the rug”（じゅうたんの下へほうきで掃く）と言うことがある。

[訳注 60] メタ分析 (meta-analysis) とは、すでに実施された研究の結果を統合して分析することを指す。

[訳注 61] CONSORT は Consolidated Standards of Reporting Trials（試験の報告の統合された標準）の略で、ランダム化比較試験である臨床試験において、どのようなことを報告しなくてはならないかについてまとめている。CONSORT 2010 声明の日本語版は以下の

を設けることで、この問題に対抗しはじめている。論文の著者には、研究内容を投稿する前にチェックリストの要求に従うことが求められている。そして、編集者には、関連する詳細の記述が全て含まれているかを確かめることが求められている。チェックリストはうまくいっているようだ。ガイドラインに従う学術誌で公刊された研究は、全ての本質的な詳細でないにせよ、より本質的な詳細を報告する傾向がある［文献 50］。それにもかかわらず、残念なことに、基準が一貫性なく適用され、しばしば詳細の記述が欠けた研究がすりぬけてしまう［文献 45］。学術誌の編集者は、報告基準を遵守させるために、より一層の努力をする必要があるだろう。

公刊された論文があまりうまくいっていないことを見てきた。公刊されていない研究についてはどうだろうか。

## 書類棚の中の科学

先に、研究結果に対する多重比較と真実の誇張の影響を見てきた。研究において、検定力の低い状態でたくさんの比較をする場合、こうした問題が発生する。そして、高い偽陽性率と誇張された効果量の推定がもたらされることになる。こうした問題は公刊された研究の至るところに見られる。

だが、全部の研究が公刊されるわけではない。例えば、医学では、「この薬を試したが、効かなかったようだ」ということをわざわざ公刊しようとする科学者はほとんどいない［訳注 62］から、医学研究のごく一部しか目にする

URL に掲載されている。
`http://www.consort-statement.org/Media/Default/Downloads/Translations/Japanese_jp/Japanese%20CONSORT%20Statement.pdf`

［訳注 62］ここでは、効かなかったものをわざわざ公刊しようとする科学者はほとんどいないと書いてあるが、科学者があえて公刊しようと思った場合はどうなるだろうか。実のところ、科学者が公刊しようと思っても、効かなかったものについての公刊を承諾する学術誌はほとんどないと考えられる。なぜかと言うと、学術誌に論文を載せる場合、その論文には、新しく分かったことで、意味があることを載せる必要がある。しかし、効かなかったという知見は、全く無意味というわけではないが、新しくて意味があることとするにはパンチが足りず、結果として公刊されないのである。

ことがない。

腫瘍抑制タンパク質の TP53 とその頭頸部ガンへの影響についての研究という事例を考えてみよう。TP53 を測定してガン死亡率を予測できるだろうということが多くの研究で示唆されている。なぜならば、TP53 は、細胞の成長と発達を調整するはたらきを持つがゆえに、ガンを防ぐために正確に機能するにちがいないからだ。TP53 とガンに関する公刊された 18 の研究全てをまとめて分析した場合[訳注 63]、統計的にかなり有意な相関が得られる。腫瘍が人を死に至らせる可能性を判断するために、TP53 を測定することができよう。

しかし、TP53 に関する公刊されていない結果——他の研究で言及されてはいるが、公刊あるいは分析されていないデータ——を発掘してみたとしよう。こうしたデータを混ぜ合わせると、統計的に有意な効果は消えてしまう[文献 39]。結局のところ、相関がないことを示すデータをわざわざ公刊しようとする著者がほとんどいないために、メタ分析においては偏った標本しか使えなかったのである。

似たような研究が、ファイザー[訳注 64]の売っているレボキセチンという抗うつ剤について調べている。いくつかの公刊された研究において、偽薬に比べてレボキセチンは効果があることが示唆されている。これによって、いくつかのヨーロッパの国では、うつ病の患者に処方することを承認している[訳注 65]。治療の評価に責任を負っているドイツの医療品質・効率性研究機構[訳注 66]は、ファイザーから公刊されていない試験データを何とか手に入れた。公刊されていないデータは公刊されていたものの 3 倍以上に及んでいた。そして、医療品質・効率性研究機構がそのデータを注意深く分析した

[訳注 63] ここでは 18 の研究を合わせたメタ分析を実施していることになる。

[訳注 64] ファイザー (Pfizer) は、1849 年に設立されたアメリカの大手製薬会社である。

[訳注 65] 日本やアメリカではレボキセチンが承認されていない。

[訳注 66] ドイツの医療品質・効率性研究機構 (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen; IQWiG) は、2004 年にドイツの医療制度改革によって設立された独立機関で、薬品などの品質や効率性について研究を実施している。

ところ、レボキセチンは効果がないことが分かった。ファイザーは効果がないと示す研究について言及しないことで、薬に効果があると大衆に説得していただけなのである［文献 18］。

この問題は公刊バイアス (publication bias) またはファイル引き出し問題 (file-drawer problem) として一般に知られている。多くの研究が、ファイルを入れる引き出しに何年も収められたままで、貢献できるかもしれない価値あるデータにもかかわらず、決して公刊されないのである。

この問題は、単に公刊された結果の偏りだけをもたらすものではない。研究が公刊されないことは、苦労を繰り返すことにつながる。もし、すでに行われた研究について知らなければ、他の科学者がもう一度その研究を行うかもしれない。そうなれば、金銭と労力の無駄である。

規制を行う側と学術誌は、この問題を止めようと試みている。アメリカの食品医薬品局［訳注 67］は、ある種の臨床試験に対して、試験を始める前に、食品医薬品局が運営しているウェブサイト ClinicalTrials.gov で登録することを求めており、さらに、試験が終わってから 1 年以内に結果を公刊することも求めている。同様に、医学誌編集者国際委員会は、2005 年に、事前に登録されていない研究については公刊しないことを表明している。

不幸なことに、738 件の登録された臨床試験に関する報告によれば、22% しか公刊するという法的要件を満たしていなかった［文献 51］。食品医薬品局は法令遵守違反で製薬会社に罰金をとることをしていない。また、学術誌は、試験登録の要求を常に強制しているわけではない。ほとんどの研究は単に消えているだけである。

［訳注 67］アメリカの食品医薬品局 (The Food and Drug Administration; FDA) は保険社会福祉省の部門の 1 つであり、食品や医薬品の安全検査、認可などを所管している。

# 11 何をしてきたか

訳者による概要説明
この章では、現在の科学研究で発生している統計分析上の問題の事例を紹介している。

私はここまで容赦ない絵を描いてきた。だけれども、公刊された研究の小さな詳細を捉えて、すさまじい量の誤りのリストを作ることは誰にでもできる。これらの問題は重要なのだろうか？

うん、そうだ。そうじゃなかったらこの文章を書いていない。

ジョン・ヨアニディスの有名な論文「なぜほとんどの公刊された研究上の発見は間違っているのか」[文献 33] は、研究結果の実証試験よりもむしろ数学上の懸念点に基づいたものである。もし、ほとんどの研究論文が検定力が不足しているのであれば――実際そうなのだが――多くの手法の中から都合の良い結果を得るための手法を選ぶ自由があるのならば――実際そうなのだが――ほとんどの検定された仮説が誤っていてほとんどの真の仮説がとても小さな効果量と対応しているのであれば、多くの偽陽性を得ることが数学的に確定している。

もし実証的な結果を知りたければ、ジョン・ヨアニディスとジョナサン・シェーンフェルドのおかげでそれを知ることができる。彼らは「我々が食べるものは何でもガンと関係するのか」という問題を研究した[原注 6][文献 55]。料理本からありふれた食材を 50 種類選んだ後、これらの食材とガン罹患率とを結びつけている研究を探すことに着手した。すると、40 種類の食材で

[原注 6] 現在継続中の腫瘍学オントロジープロジェクト (Oncological Ontology Project) の重要な部分は、全てのものをガンを治すものとガンを起こすものの 2 種類に分けるものである。

216 個の研究を発見した。もちろん、研究のほとんどが互いに一致していなかった。ほとんどの食品について、ガンになるリスクを増加させると主張する研究と減少させると主張する研究の両方があった。ほとんどの統計的な証拠は弱いもので、メタ分析からは大概元の研究よりずっと小さな効果しかないことが示された。

もちろん、追跡研究やメタ分析で矛盾が起きていることは、論文が正しいものかのように引用されることを妨げない。明白な結果がある大量の追跡試験と矛盾している効果についても、5 年あるいは 10 年後にしばしば引用され、科学者がその結果が誤っていると気づいていないことがある［文献 60］。もちろん、新しい発見というものは広くマスメディアに報道されるものであるのに対し、矛盾や修正というのはほとんど言及されない［文献 23］。科学者が知らなかったとしてもその科学者を非難しがたいのだ。

単なる偏った結果を忘れないようにしよう。医学誌における低劣な報告基準は、統合失調症の新しい治療法を試す研究で、症状を評価するのに使った尺度について報告することをしないで済ませうることを意味する。偏りはここから手軽に生まれる。公刊されていない尺度を用いた試験は、かつて有効だと検証された試験を用いるよりも良い結果を生み出しがちであるからだ［文献 43］。他の医学研究では特定の結果が不都合だったりつまらなかったりしたら、単純にそれを除外している。このことにより、その後のメタ分析では前向きな結果しか含まれなくなるという偏りが生じてしまう。メタ分析の 3 分の 1 がこの問題によって悪影響を受けていると推定されている［文献 37］。

医学の至適基準［訳注 68］を考慮して、メタ分析とその後の大規模なランダム化比較試験とを比較した他の調査によると、3 分の 1 以上の事例でランダム化された試験の結果はメタ分析の結果とうまく合わなかった［文献 42］。他のメタ分析とその後の研究の比較では、ほとんどの結果が誇張されていることと、おそらく 5 分の 1 が偽陽性であることが示されている［文献 49］。

［訳注 68］医学において、至適基準 (gold standard) とは、診断をする時に、最も正確に診断ができる検査方法のことを指す。

信頼区間を誤って使っている多数の自然科学の論文のことを忘れないようにしよう［文献 40］。あるいは探索的研究で統制されていない多重比較をもとに念力の証拠を挙げていることになっている査読された心理学の論文を忘れないようにしよう［文献 63］。当然のことながら、結果を再現することに失敗する。検定において検定力を計算していないと思われる科学者によって［文献 20］。

我々は問題を抱えている。改善に取り組もう。

# 12 何ができるだろうか

訳者による概要説明
この章では、統計の誤りを防ぐために、統計教育を充実させることと、学術誌が努力すべきであることが重要であると提言している。

このガイドを通じてたくさんの統計に関する問題について論じてきた。こうした問題は、医学、物理学、気候科学、生物学、化学、神経科学などの科学の多くの分野で見られる。統計的手法を使ってデータを分析しようとする研究者は誰でも誤りを犯しうる。そして、今まで見てきたように、ほとんどの人が誤りを犯している。このことに対して何ができるだろうか？

## 統計教育

アメリカの科学の学生のほとんどが、最低限度の統計教育しか受けていない。多くの学生は、もしかしたら必修が 1、2 科目といったところかもしれないし、あるいは全く受けていないかもしれない。そして、たとえ学生が統計の授業を取ったとしても、学生は適切な統計技法を完全に理解することは決してなく（あるいは単に技法を忘れてしまって）、科学上の問題に対して統計的概念を適用することができないと大学の教員は言っている。この状況は変える必要がある。ほとんど全ての科学に関する学問分野は実験データの統計分析次第のものだ。統計上の誤りは研究助成金や研究者の時間を無駄にするのだ。

いくつかの大学では、学生が専門分野の問題に対してすぐに統計知識を適用できるように、統計の授業を科学の授業に統合する実験的試みを行っている。予備的な結果によると、これらの手法はうまくいっている。学生は統計

についてより多く学び、それを忘れないようになった。そして、強制的に統計の授業を取らされることに対して学生がぶつくさ文句を言う時間は減った[文献 44]。より多くの大学が、どの手法が最もうまくはたらくかを見るためにコンセプトテストを用いた上で、このような手法を採用すべきである。

もっと自由に手に入る教材も必要だ。私は、実験室でのデータを分析する必要があって、どうすれば分からなかった時に、統計に初めて触れた。しっかりとした統計教育がもっと広まるまでは、多くの学生が私と同じような状態になって、教材が必要になるだろう。OpenIntro Stats[訳注 69]のようなプロジェクトは有望であり、近い将来より多くのものが見られると期待している。

## 科学に関する出版

学術誌は、私が論じた問題の多くを解決するためにゆっくりと前進している。ランダム化試験のための CONSORT のような、報告に関するガイドラインは、公刊される論文が再現可能であるようにするために必要となる情報が何かということをはっきりさせている。残念なことに、今まで見てきたように、これらのガイドラインが強制されることは少ない。私たちは、著者がより厳格な基準を守るように、学術誌に圧力をかけ続けなくてはならない。

一流の学術誌はこの先頭に立つべきである。『ネイチャー』はそうし始めており、記事が出版される前に著者が埋める必要がある新しいチェックリストを発表している。このチェックリストでは、標本の大きさ、検定力の計算、臨床試験登録番号、完全な CONSORT チェックリスト、多重比較をするための調整について報告すること、そしてデータとソースコードを共有することを求めている。このガイドラインは『ダメな統計』に示されているほとんどの問題を対象としている。『ネイチャー』は、必要がある時に論文に

[訳注 69] OpenIntro Statistics はインターネット上で公開されている自由に使える統計の教材である。入門から始まり、かなり高度な話題まで載っている。ただし、易しくはなく、他の入門教材に比べるとかなり難しい部類に入る。

ついて統計学者に相談できるようにしている。

これらのガイドラインが強制されれば、結果はずっと信頼でき、再現可能な科学研究となるだろう。もっと多くの学術誌が同じようにすべきだ。

## あなたがすべきこと

あなたがすべきことは、以下の単純な 4 つのステップで表現することができる。

1. 統計の教科書を読むか、良い統計の授業を取れ。練習せよ。
2. 学んだ誤解や誤りを避けるために、自分のデータ分析について、注意深く慎重に計画せよ。
3. もし、$p$ 値の単純な誤解のような、ありふれた誤りを科学の文献で見つけたら、犯人の頭を統計学の教科書でどつけ[訳注 70]。これは治療だ。
4. 科学教育と科学出版の変化を求めよ。これは我々の研究だ。台無しにするな。

[訳注 70] 実際に教科書でどついた場合にもたらされる結果について、訳者は責任を負いかねる。

❖ 訳者コラム：ウェブ上で公開されている統計入門教材 ❖

本文中で紹介されていた OpenIntro Statistics のほかにも、ウェブ上で公開されている統計入門教材には様々なものがある。その中からいくつかの教材を紹介しよう。

基礎からの統計学　内容は、記述統計、確率、確率分布、簡単な推定・検定、相関と線形回帰などごく基本的なもの。Excel での計算方法も掲載。http://www.heisei-u.ac.jp/ba/fukui/text.html にて閲覧できる。

統計学入門　扱っている内容は、上述の『基礎からの統計学』とほぼ同じ。理論的な話が多く、実践的な話がやや少ないのが玉に瑕(きず)だ。http://ruby.kyoto-wu.ac.jp/~konami/Text/ にて閲覧できる。

gacco　JMOOC が公認するオンライン授業配信サイトの gacco (http://gacco.org/) でも、日本語で統計学の入門授業が開講されている。

Collaborative Statistics　英語。内容の面でも英語の面でもかなり易しく書かれており、初学者にはおすすめ。分量は多いが、その分だけ内容も豊富。練習問題が事細かく書いてあることもうれしい。http://cnx.org/content/col10522/latest/ にて閲覧できる。

Online Statistics　英語。これも統計に関する入門教材となっているウェブサイトである。http://onlinestatbook.com/ にて閲覧できる。

# 13 終わりに

訳者による概要説明
この章は、最後の章として本書全体の内容を簡潔にまとめたものである。

間違った確信に注意せよ。そのうち、他の人のようなへまを自分はしないという自己満足におちいってしまうかもしれない。私はデータ分析の数学的処理に関するしっかりとした入門は教えていない。概念に関する単純な誤りを越えて、統計で失敗する方法はたくさんあるのだ。

誤りはしばしば起きるだろう。なぜならば、どういうわけか、科学の学部課程や医学校[訳注 71]で、統計と実験デザインに関する授業を必修とするものはほとんどないからだ。そして、統計学の入門授業では、検定力と多重推論[訳注 72]についての問題をとばしてしまうこともある。現代の科学の営みにおいてデータと統計分析が最も重要な役割を果たしているにもかかわらず、こうした状況が容認されている。薬の処方の経験がない医者を容認することはないはずだ。だとすれば、なぜ統計の訓練を受けていない科学者を容認するのだろうか。科学者には統計に関する正式な訓練とアドバイスが必要だ。

[訳注 71] 日本では医師養成は学部課程で行われるが、米国では医学校 (medical school) という専門大学院で医師が養成される。

[訳注 72] 多重推論 (multiple inference) とは同じデータに対して、複数回の統計的推論を行ってしまうことを指す。

> 実験が終わった後に統計学者に相談することは、しばしば単に検死を頼むようなものになる。統計学者は、何のせいで実験が死んだのかについて言うことができるかもしれない。
>
> ——$p$ 値を普及させた人、R. A. フィッシャー[訳注 73]

学術誌は、質の低い統計分析を行っている研究を却下することを選択してもよい。新しいガイドラインとプロトコルは、いくつかの問題を消し去るだろう。しかし、統計の原理について十分に訓練を受けた科学者が出てこないかぎり、実験デザインとデータ分析は改善しないだろう。統計的な有意差を全力で求めることが続くだけだ。

変化は簡単ではないだろう。厳格な統計基準はただではやってこないのだ。例えば、もし科学者が検定力の計算を日常的に行うようになったら、確かな結論に至るにははるかに多くの標本サイズが必要になることにすぐに気づくだろう。臨床試験はただではない。そして、研究に費用がよりかかることは、公刊される試験がより少なくなることを意味する。必要もないのに科学の進歩が遅くなってしまうことに反対するかもしれない。だが、信頼できない結果に基づいて進歩することはもっとひどいことではないだろうか。

科学の学生の皆様へ。機会があれば、1 科目か 2 科目、統計の授業に投資してください。研究者の皆様へ。訓練、良い書籍、統計に関するアドバイスに投資してください。そして、お願いですから、次に誰かが「この結果は $p < 0.05$ で有意だから、これが偶然である確率は 20 分の 1 しかない！」と言うのを聞くことがあったら、私のためにその連中の頭を統計の教科書でぶったたいてください[訳注 74]。お願いです。

---

[訳注 73] ロナルド・エイルマー・フィッシャー (Ronald Aylmer Fisher, 1890-1962) は 20 世紀の最も偉大な統計学者の 1 人で、遺伝学の研究者としても知られている。フィッシャーが成し遂げた統計学上の業績には様々なものがあるが、中でも分散分析 (ANOVA) や実験計画法を発展させたことが重要な業績として挙げられる。

[訳注 74] 実際に教科書でぶったたいてどんな結果がもたらされたとしても、訳者は責任を負いかねるので、読者諸氏は注意されたい。

お断り：このガイドでの助言は、訓練された統計の専門家の助言に換えられるものではありません。もし、あなたが深刻な統計上の誤りに悩まされていると思っているようでしたら、直ちに統計学者に相談してください。あなたがこのガイドを使用した結果として、あなたの名誉が傷ついたり、統計に関する過誤や誤解が起きたりしても、それらに対して私は責任を負いかねます。

証拠を詳細に検討することなしに、科学研究の結果を却下することの正当化のためにこのガイドを使うことは、非常に大型の統計の教科書で頭のてっぺんをバシバシたたく理由となりえます[訳注 75]。このガイドは、統計的な誤りを見つけるのを助ける存在であるべきで、嫌いな科学を選んで無視することを許す存在ではありません。

[訳注 75] 似たような話の繰り返しとなるが、実際に教科書でバシバシたたいた後にもたらされる結果について、訳者は責任を負いかねる。

# 参考文献

[1] K. A. Baggerly, K. R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, 3:1309—1334, 2009.

[2] M. Bakker, J. M. Wicherts. The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43:666—678, 2011.

[3] D. Bassler, M. Briel, V. M. Montori, M. Lane, P. Glasziou, Q. Zhou, D. Heels-Ansdell, S. D. Walter, G. H. Guyatt. Stopping randomized trials early for benefit and estimation of treatment effects: Systematic review and meta-regression analysis. *JAMA*, 303:1180—1187, 2010.

[4] P. L. Bedard, M. K. Krzyzanowska, M. Pintilie, I. F. Tannock. Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology Annual Meetings. *Journal of Clinical Oncology*, 25:3482—3487, 2007.

[5] C. G. Begley, L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483:531—533, 2012.

[6] S. Belia, F. Fidler, J. Williams, G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10:389—396, 2005.

[7] Y. Benjamini, Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 289—300, 1995.

[8] C. Bennett, A. Baird, M. Miller, G. Wolford. Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results*, 1:1—5, 2010.

[9] A. F. Bogaert. Biological versus nonbiological older brothers and

men's sexual orientation. *PNAS*, 103:10771—10774, 2006.

[10] R. Bramwell, H. West. Health professionals' and service users' interpretation of screening test results: Experimental study. *BMJ*, 2006.

[11] C. G. Brown, G. D. Kelen, J. J. Ashton, H. A. Werman. The beta error and sample size determination in clinical trials in emergency medicine. *Annals of Emergency Medicine*, 16:183—187, 1987.

[12] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, M. R. Munafò. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 2013.

[13] J. Carp. The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage*, 63:289—300, 2012.

[14] A. Chan, A. Hróbjartsson, M. T. Haahr, P. C. Gøtzsche, D. G. Altman. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, 291:2457—2465, 2004.

[15] A. Chan, A. Hróbjartsson, K. J. Jørgensen, P. C. Gøtzsche, D. G. Altman. Discrepancies in sample size calculations and data analyses reported in randomised trials: Comparison of publications with protocols. *BMJ*, 337:a2299, 2008.

[16] K. C. Chung, L. K. Kalliainen, R. A. Hayward. Type II (beta) errors in the hand literature: The importance of power. *The Journal of Hand Surgery*, 23:20—25, 1998.

[17] H. H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12:335—359, 1973.

[18] D. Eyding, M. Lelgemann, U. Grouven, M. Härter, M. Kromp, T. Kaiser, M. F. Kerekes, M. Gerken, B. Wieseler. Reboxetine for acute treatment of major depression: Systematic review and meta-analysis

of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ*, 341:2010.

[19] K. R. Gabriel. A simple method of multiple comparisons of means. *Journal of the American Statistical Association*, 73:724—729, 1978.

[20] J. Galak, R. A. LeBoeuf, L. D. Nelson, J. P. Simmons. Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, 103:933—948, 2012.

[21] A. Gelman, P. Price. All maps of parameter estimates are misleading. *Statistics in Medicine*, 18:3221—3234, 1999.

[22] A. Gelman, H. Stern. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, 60:328—331, 2006.

[23] F. Gonon, J. P. Konsman, D. Cohen, T. Boraud. Why most biomedical findings echoed by newspapers turn out to be false: The case of attention deficit hyperactivity disorder. *PLoS ONE*, 7:e44275, 2012.

[24] S. N. Goodman. Toward evidence-based medical statistics. 1: The *p* value fallacy. *Annals of Internal Medicine*, 130:995—1004, 1999.

[25] S. N. Goodman. A dirty dozen: Twelve *p*-value misconceptions. *Seminars in Hematology*, 45:135—140, 2008.

[26] P. C. Gøtzsche. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials*, 10:31—56, 1989.

[27] P. C. Gøtzsche. Believability of relative risks and odds ratios in abstracts: Cross sectional study. *BMJ*, 333:231—234, 2006.

[28] A. Grafen, R. Hails. *Modern Statistics for the Life Sciences*. Oxford, England: Oxford Univ. Press, 2002.

[29] E. Hauer. The harm done by tests of significance. *Accident Analysis & Prevention*, 36:495—500, 2004.

[30] D. Hemenway. Survey research and self-defense gun use: An ex-

planation of extreme overestimates. *The Journal of Criminal Law and Criminology*, 87:1430—1445, 1997.

[31] K. Huwiler-Müntener, P. Jüni, C. Junker, M. Egger. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA*, 287:2801—2804, 2002.

[32] J. P. A. Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294:218—228, 2005.

[33] J. P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2:e124, 2005.

[34] J. P. A. Ioannidis. Why most discovered true associations are inflated. *Epidemiology*, 19:640—648, 2008.

[35] J. P. A. Ioannidis, T. A. Trikalinos. Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, 58:543—549, 2005.

[36] V. E. Johnson. Revised standards for statistical evidence. *PNAS*, 110:19313—19317, 2013.

[37] J. J. Kirkham, K. M. Dwan, D. G. Altman, C. Gamble, S. Dodd, R. Smyth, P. R. Williamson. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*, 340:c365, 2010.

[38] W. Krämer, G. Gigerenzer. How to confuse with statistics or: The use and misuse of conditional probabilities. *Statistical Science*, 20:223—230, 2005.

[39] P. A. Kyzas, K. T. Loizou, J. P. A. Ioannidis. Selective reporting biases in cancer prognostic factor studies. *Journal of the National Cancer Institute*, 97:1043—1055, 2005.

[40] J. R. Lanzante. A cautionary note on the use of error bars. *Journal of climate*, 18:3699—3703, 2005.

[41] S. E. Lazic. The problem of pseudoreplication in neuroscientific studies: Is it affecting your analysis?. *BMC Neuroscience*, 11:5, 2010.

[42] J. LeLorier, G. Gregoire, A. Benhaddad. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, 1997.

[43] M. Marshall, A. Lockwood, C. Bradley, C. Adams, C. Joy, M. Fenton. Unpublished rating scales: A major source of bias in randomised controlled trials of treatments for schizophrenia. *The British Journal of Psychiatry*, 176:249—252, 2000.

[44] A. M. Metz. Teaching statistics in biology: Using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. *CBE Life Sciences Education*, 7:317—326, 2008.

[45] E. Mills, P. Wu, J. Gagnier, D. Heels-Ansdell, V. M. Montori. An analysis of general medical and specialist journals that endorse CONSORT found that reporting was not enforced consistently. *Journal of Clinical Epidemiology*, 58:662—667, 2005.

[46] D. Moher, C. S. Dulberg, G. A. Wells. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*, 272:122-124, 1994.

[47] V. M. Montori, P. J. Devereaux, N. Adhikari. Randomized trials stopped early for benefit: A systematic review. *JAMA*, 294:2203—2209, 2005.

[48] S. Nieuwenhuis, B. U. Forstmann, E. Wagenmakers. Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, 14:1105—1109, 2011.

[49] T. V. Pereira, J. P. A. Ioannidis. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology*, 64:1060—1069, 2011.

[50] A. C. Plint, D. Moher, A. Morrison, K. Schulz, D. G. Altman, C. Hill,

I. Gaboury. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical journal of Australia*, 185:263—267, 2006.

[51] A. P. Prayle, M. N. Hurley, A. R. Smyth. Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: Cross sectional study. *BMJ*, 344:d7373, 2011.

[52] D. F. Preusser, W. A. Leaf, K. B. DeBartolo, R. D. Blomberg, M. M. Levy. The effect of right-turn-on-red on pedestrian and bicyclist accidents. *Journal of Safety Research*, 13:45—55, 1982.

[53] F. Prinz, T. Schlange, K. Asadullah. Believe it or not: How much can we rely on published data on potential drug targets?. *Nature Reviews Drug Discovery*, 10:328—329, 2011.

[54] N. Schenker, J. F. Gentleman. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55:182—186, 2001.

[55] J. D. Schoenfeld, J. P. A. Ioannidis. Is everything we eat associated with cancer? A systematic cookbook review. *American Journal of Clinical Nutrition*, 97:127—134, 2013.

[56] S. Schroter, N. Black, S. Evans, F. Godlee, L. Osorio, R. Smith. What errors do peer reviewers detect, and does training improve their ability to detect them?. *JRSM*, 101:507—514, 2008.

[57] J. P. Simmons, L. D. Nelson, U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22:1359—1366, 2011.

[58] D. G. Smith, J. Clemens, W. Crede, M. Harvey, E. J. Gracely. Impact of multiple comparisons in randomized clinical trials. *The American Journal of Medicine*, 83:545—550, 1987.

[59] J. M. Steele. Darrell Huff and fifty years of *How to Lie with Statistics*.

*Statistical Science*, 20:205—209, 2005.

[60] A. Tatsioni, N. G. Bonitsis, J. P. A. Ioannidis. Persistence of contradicted claims in the literature. *JAMA*, 298:2517—2526, 2007.

[61] S. Todd, A. Whitehead, N. Stallard, J. Whitehead. Interim analyses and sequential designs in phase III studies. *British Journal of Clinical Pharmacology*, 51:394—399, 2001.

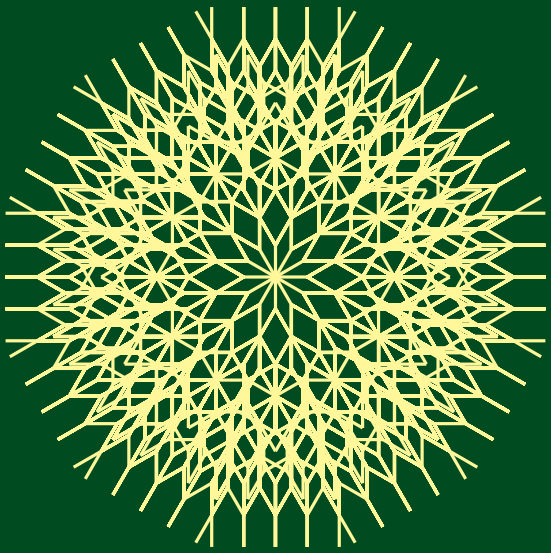
[62] R. Tsang, L. Colley, L. D. Lynd. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *Journal of Clinical Epidemiology*, 62:609—616, 2009.

[63] E. Wagenmakers, R. Wetzels. Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 2011.

[64] H. Wainer. The most dangerous equation. *American Scientist*, 95:249—256, 2007.

[65] J. M. Wicherts, M. Bakker, D. Molenaar. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, 6:e26828, 2011.

[66] J. M. Wicherts, D. Borsboom, J. Kats, D. Molenaar. The poor availability of psychological research data for reanalysis. *American Psychologist*, 61:726—728, 2006.



〔統計に関する〕誤りはしばしば起きるだろう。なぜならば、どういうわけか、科学の学部課程や医学校で、統計と実験デザインに関する授業を必修とするものはほとんどないからだ。そして、統計学の入門授業では、検定力と多重推論についての問題をとばしてしまうこともある。現代の科学の営みにおいてデータと統計分析が最も重要な役割を果たしているにもかかわらず、こうした状況が容認されている。薬の処方の経験がない医者を容認することはないはずだ。だとすれば、なぜ統計の訓練を受けていない科学者を容認するのだろうか。科学者には統計に関する正式な訓練とアドバイスが必要だ。

――「終わりに」より