ソース: https://biostatistics.ucdavis.edu/sites/g/files/dgvnsk4966/files/inline-files/Greenland.Advancing%20statistics%20reform%2C%20part%204.Slides%201-110%2C%2001%20June%202022.pdf (https://biostatistics.ucdavis.edu/sites/g/files/dgvnsk4966/files/inline-files/Greenland.Advancing%20statistics%20reform%2C%20part%204.Slides%201-110%2C%2001%20June%202022.pdf)

統計学の改革の推進:
抵抗への直面の中で思考と実践を改善する方法
以下からの発展:
「合理的な認知に対する統計学の専制の打破」
「確率と推論より認知と因果を優先」
「統計で嘘をつかない方法:
推論ではなく記述的なイメージを用いよ」
「統計学の教育と実践における認知科学と因果性の必要性」
「統計は取り壊し予定の建物: 解体と再建の計画」

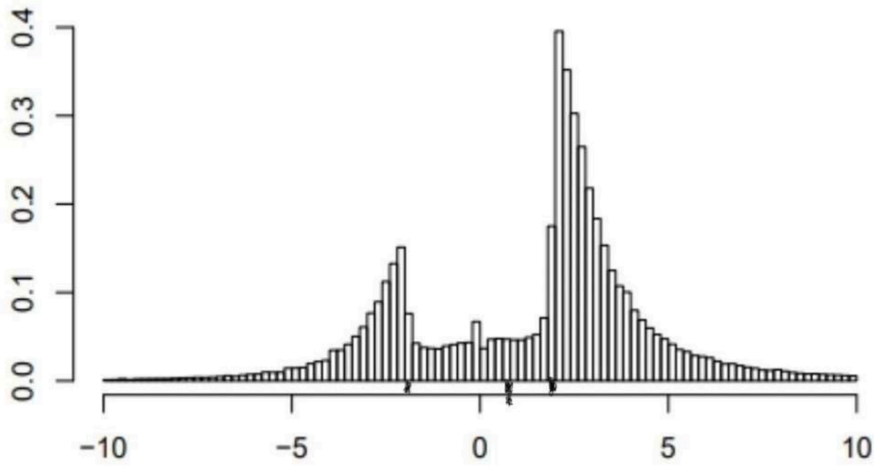
Sander Greenland, Dept of Epidemiology and Dept of Statistics, UCLA
Please report errors and send comments to Sander Greenland at lesdomes@ucla.edu

---

科学は葬式ごとに進歩するが、
統計学において権威は不滅である

- **英雄的なストーリー:** 科学は、各世代が先人の考えに挑戦し、厳しい**実証的(empirical)**検証に耐えられなかったものを捨て去ることによって進歩する。
- 対照的に、**統計学は、伝統的な方法論を神聖化**し、学術的な数学的・哲学的訴えによってそれらを擁護し、公衆への情報提供に対する害を軽視**することによって衰退してきた。**

帰結: van Zwet & Cator 2021 図1
1976年から2019年のMedlineから得られた100万件以上のz値。
推定された分布曲線は右に歪み、75%以上が0より大きい。



訳註: 上の図は次のオープンアクセス論文の図1より:

- Erik W. van Zwet, Eric A. Cator, The significance filter, the winner's curse and the need to shrink, 2021, https://doi.org/10.1111/stan.12241 (https://doi.org/10.1111/stan.12241)

- **ゼロ主義(nullism)や二分法への執着(dichotomania)のような認知バイアス達を「科学的原則」として神聖化し、数学的枠組みをあたかも物理的現実であるかのように扱い(現実との混同(reification))、確実性や最終性を渇望する人間のバイアスを無視することが、統計学の中核を腐敗させてきた。**

- **解決策: 統計学を、確率論の一分野としてではなく、認知科学と因果論を中核構成要素とする情報科学として再構築する。**

訳註: 以下のように訳した:

- nullism→ゼロ主義
- dichotomania→二分法への執着
- reification→現実との混同

“reification”は「抽象的なものを具体的なものとして扱うこと」という意味で「実体化」「具象化」と訳されることが多いが、この翻訳では分かり易さのために「現実との混同」と訳すことにした。

デ・フィネッティ(De Finetti)の急進的ベイズ主義では、すべての確率は「主観的」であり、観察者の心の性質だけを記述する。その見方では

- パターンが「偶然によって引き起こされる」という考えは、世界についての因果的言明としては馬鹿げている。
- むしろ我々は、認識されたパターンの**因果的説明**を求める際に、もっともらしいと提案された少数の因果的可能性達の中からの全然無作為でない(非常に偏った)選択について考慮する。
- そして我々は、考慮されなかった因果的説明達の残余無限性(residual infinitude)を、「偶然」と呼ばれる形而上学的な原因を形成するものとして、現実と混同(reify)してしまう。

訳註:

- ブルーノ・デ・フィネッティ - Wikipedia (https://ja.wikipedia.org/wiki/%E3%83%96%E3%83%AB%E3%83%BC%E3%83%8E%E3%83%(Bruno de Finetti - Wikipedia (https://en.wikipedia.org/wiki/Bruno_de_Finetti))
- “reify”は「抽象的なものを具体的なものとして扱う」という意味で、「実体化する」「具象化する」と訳されることが多いが、この翻訳では分かり易さの重視のために「現実と混同する」と訳すことにした。

- **すべての分析は、著しく不完全な感度分析の一部と見なされるべきである。**
- **頻度論 vs ベイズの論争は、詳細な論理的分析の下では消え去る宗教論争である。**
- ボックスの見解(Boxian view): ベイズ的なツールは方法とモデル構築のために用いられ、頻度論のツールはそれらの評価のために用いられる(他にも多くの有用な組み合わせがある)。
- つい最近まで、両方の「学派」は自らの手法が導かれるべき本質的な因果的/文脈的次元をカバーできていなかった。

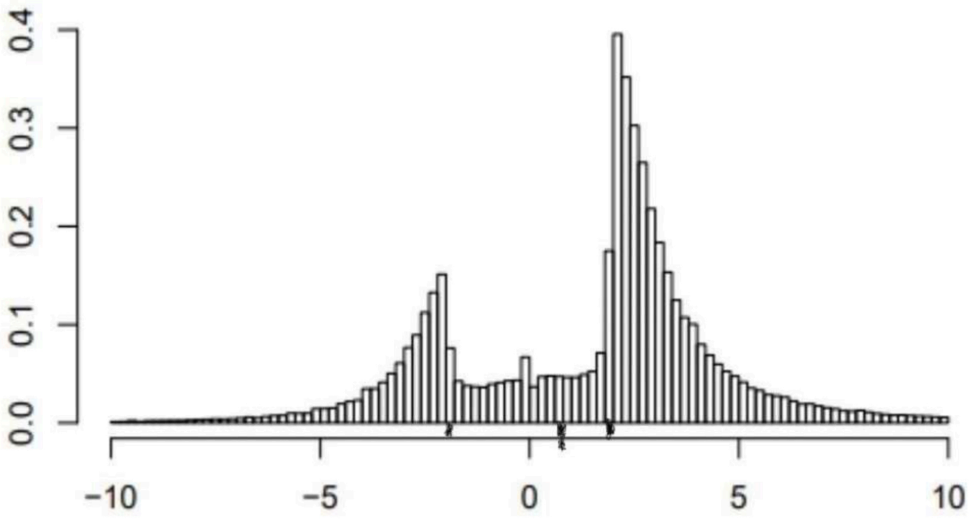
訳註: ボックスとボックスの見解については以下を参照せよ:

- George E. P. Box - Wikipedia (https://en.wikipedia.org/wiki/George_E._P._Box)
- George E. P. Box, Sampling and Bayes’ Inference in Scientific Modelling and Robustness, Royal Statistical Society. Journal. Series A: General, Volume 143, Issue 4, July 1980, Pages 383–404, https://doi.org/10.2307/2982063 (https://doi.org/10.2307/2982063), Free (https://academic.oup.com/jrsssa/article/143/4/383/7105478), Google Scholar (https://scholar.google.co.jp/scholar?cluster=5019602469235484232)

これらの推論概念を統一するものは何か？
— 確率ではなく因果である:

- 過去の原因: **我々の観察の原因となったもの (「説明」するもの) は何か？**これは**物理的メカニズム**に関する問いであり、確率のようなその振る舞いの抽象化ではない。
- 未来の効果: **行動は未来にどう影響するか？**これは、実際のイベント頻度のようなメカニズムの振る舞いをどう変えるかに関する問いであり、確率分布に関する問い**ではない。**
- 例: 改革の効果はどうなるか？…

解答: 研究結果に基づく選択的報告が
生じ続けるような**いかなる**改革も、
全体との比較で利用可能な結果の分布を歪めてしまう。



推論(INFERENCE)とは何か？

- 辞書の例: **「証拠と論理的思考に基づいて到達した結論。」**
- **科学的推論とは現実に関する複雑だが厳密に制御された判断であり、以下の仮定に基づいている:**
  論理的に一貫した「客観的」(観察者の外部にある)現実が存在し、それが発見可能な法則に従って我々の知覚を引き起こす:
  私の知覚 ← 現実 → あなたの知覚

* **これによって推論は認知科学の一部として位置づけられる。**

科学的推論と「統計的推論」の対比

- すべての形式化(formalisms)、すべての「学派(schools)」、すべてのツールキット(toolkits)において、「**統計的推論**」は、**データ処理プログラム(学習アルゴリズム)からの出力を受け取って、文脈から切り離されたルールを介して「推論」を生成するものになってしまっている。**
- それは、データを生成するメカニズム(データの**原因**)の過度に単純化されたモデルを抽象的な確率分布に変換する。
- **それが残す意味論的な空白は、推論上の誤りを引き起こし、(自己または他者に対する)欺瞞を助長する。**

- **統計学は意味論や日常言語を無視または軽蔑し、その代わりに「有意性(significance)」や「信頼(confidence)」を保証する欺瞞的なジャーゴンを好んだ。研究が信仰の巨大跳躍抜きでそれらに近いものを何も提供しない場合でさえそうした。**
- こうしたことは、緻密な形式化や記法や**見せかけの精度 (artificial precision)**に基づく技術的な製品やサービスを売り込むために行われた。その前提と危険性は「ソフトサイエンス」における大多数の利用者や消費者達には十分に理解されていない。
  — 医療製品の販売との類似性に注目！

訳注: 「ジャーゴン(jargon)」は「専門知識を持つ人達が好んで使う特殊な用語」というような意味である。否定的な意味で使われる場合には「業界用語」「専門バカ的な言い回し」のような意味になる。

科学界は統計科学の堕落に熱心に貢献した

自動化された狭い環境で成功したように見えたルールが、教育と研究において破壊的なフィードバックループを誘発した:

- 学生は、正解を保証するために暗記用の明確な実践ルールを求める。
- 教員は、採点の容易さを求める。
- 研究者は、受理されうる報告書を提出するためのルールを求める。
- 査読者や編集者は、査読と出版の意思決定の容易さを求める。

蔓延したルールは、強制された二分法を通じて、
とりわけ広く受け入れられて破壊的になった

- **二分法は、決定的な結論を求める人間の欲求を満たす。**なぜなら、たとえ研究が(現実の物理的なデータ生成器が) 批判的に精査されればそのような結論を強制することができなくても、二分法が適用されるからである。*

*「さらなる研究が必要」と結論した場合を除く。しかし、費用対効果や他の研究を考慮すると、それさえ正当化されないことが多い。

統計科学が数学的な骨組みの集まりに堕落した結果、
科学的推論の本質的構成要素の
詳説(explication)や訓練が置き去りにされた:

- **(確率ではなく)因果ネットワークが、データ、推論、意思決定をどのように生み出すか。**
- **認知バイアスや手続き上の問題がそれらの因果ネットワークにどのように入り込むか。**
- **価値評価** (動機、目標、実際の費用と便益) **がどのように認知に影響してすべての方法論に暗黙的に含まれるか。**

- **醜い事実: P値の主な問題は、どの統計量にも拡張される。なぜならそれらは、P値自体からではなく、真実を歪める(道理に反する)インセンティブと認知バイアスから生じるからである。**
- **道理に反するインセンティブは認知バイアス (希望的観測、心の射影) を生み出し、インセンティブが命じるものを見せる。これらのバイアスは医学のような分野の報告に蔓延している。**
- **現在、物事の見方は肯定的な報告に関するインセンティブを見るように操作されており、否定的な報告に関するインセンティブは無視されている…**

訳註: “perverse”をここでは「道理に反する」と訳した。“perverse”は「逆効果の」「偏屈な」「倒錯した」などの意味を持つ。

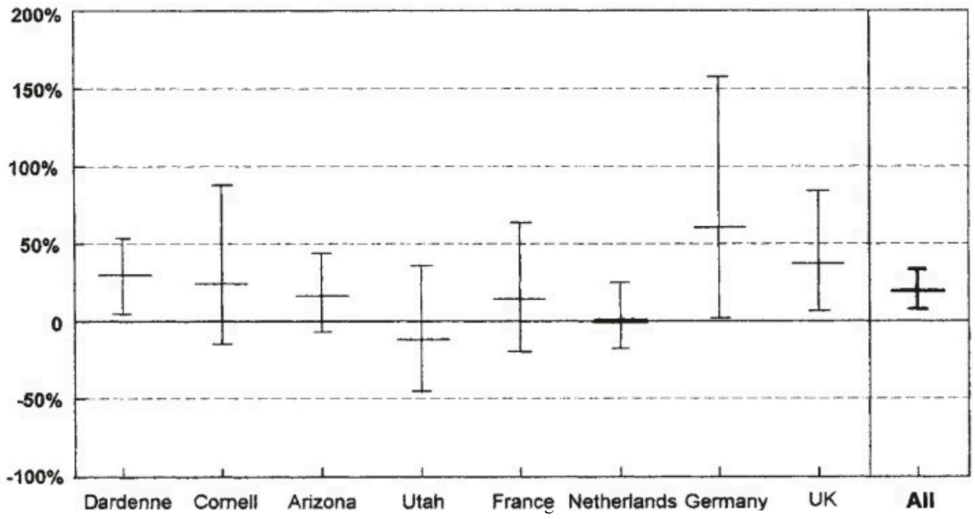
- **過去の教育や過去の実践や金銭的な利害関係への強いこだわり(commitment)に動機付けられた論理的思考(reasoning)が、真剣な改革への抵抗を駆り立てる。**

例 – 乳製品によくある表示:
「*rBSTを投与されていない牛からの牛乳。
*rBSTで処理された牛から得られた牛乳と、処理されていない牛から得られた牛乳との間に**有意差は得られていない**」

— ここでは、特定の利害関係グループが、販売に有利にするために、事実の記述に誤解を招く技術的な主張を添えることを強制している。

訳註: rBSTは遺伝子組換えウシ成長ホルモンのこと。rbSTの投与の有無についての「**有意差は得られていない**」という但し書きの強制によって特定の利害関係グループが利益を得た

Millston et al., Nature 1994: 8つの試験、rBSTで処理された牛のミルクの体細胞数(膿)が平均19%増加(メタアナリシスのP値はp=0.004):



訳註:

- E. Millstone, E. Brunner, I. White, Plagiarism or protecting public health?, Nature 1994 Oct 20;371(6499):647-8. https://doi.org/10.1038/371647a0 (https://doi.org/10.1038/371647a0)

- 「再現性の危機」は常に、出版バイアスを生み出す「統計的有意性」を探し出して発見をでっち上げようとする道理に反するインセンティブの一つとして描かれ続けている。
- **有意性の閾値を下げてもバイアスを増大させるだけである。**
- 結果に基づく**あらゆる**選択的な出版は、完全で偏りのない公的データのリポジトリを構築するという目標を損なう。
- **それでも、有意性による選択の擁護と促進は衰えることなく続いている…**

より巧妙なことに、標準的な「再現性の危機」のストーリーは、**否定的な**結果を見つけて報告するための倒錯したインセンティブの事例を無視している(例: P値を**上方にハッキングしたり、曖昧な結果を否定的と誤って報告する**ことによって)。例えば、

- 研究者、スポンサー、編集者が望ましくない関連を退けたいとき。
- 「再現の失敗」や関連への他の異議が単なる再現よりも出版されやすいとき。
- あるいはその両方…

典型的な例 (Brown et al., “Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children” (妊娠中のセロトニン作動性抗うつ薬[SSRI]使用と子供の自閉症スペクトラム障害[ASD]との関連), JAMA 2017;317:1544-52)、要旨:

- 「[Coxモデル]調整済みHR、**1.59** [95% CI, **1.17, 2.17**]。IPTW HDPS後、関連は有意ではなかった(HR, **1.61** [95% CI: **0.997, 2.59**])。」(**p = 0.0505**)
- 彼らの結論:「**子宮内での曝露は、自閉症スペクトラム障害と関連していなかった**」
- 彼らの以前のメタ分析では**HR 1.7 [1.1, 2.6]**を得ていた。

訳註:

- Brown, Hilary K. et al., Association Between Serotonergic Antidepressant Use During Pregnancy and Autism Spectrum Disorder in Children, JAMA 2017 Apr 18;

317(15):1544-1552. https://doi.org/10.1001/jama.2017.3415 (https://doi.org/10.1001/jama.2017.3415)

- IPTW (Inverse Probability of Treatment Weighting)：治療割り当ての逆確率で重み付けして交絡を調整する方法。
- HDPS (High-Dimensional Propensity Score)：高次元の交絡因子を用いた傾向スコア補正。
- HR (Hazard Ratio)：ハザード比。1より大きければ曝露がリスク上昇に関連することを示唆。
- p = 0.0505：有意差とみなすことが多い閾値 α = 0.05 をわずかに超えている。

この種の誤報を非難する論文は、少なくともKarl Pearson **1906**まで遡る:

- 「標本の大きさに対する有意性の欠如は、不注意な読者によって違いの完全な否定だとあまりにも頻繁に解釈され、これは悲惨な結果を招くかもしれない。」

この警告はその後一世紀以上にわたり数えきれないほど多くの人達によって繰り返されてきた…

それなのになぜ、こんなにもあからさまな形で続いているのか？ 単なる無知なのか？ いいえ、私は**訴訟から産業を守るために著者に強制されているのだ**と断定する。

訳注: 文献

- Karl Pearson, F.R.S., V. Note on the Significant or Non-significant character of a Sub-sample drawn from a Sample, Biometrika, Volume 5, Issue 1-2, October 1906, Pages 181–183, https://doi.org/10.1093/biomet/5.1-2.181 (https://doi.org/10.1093/biomet/5.1-2.181)

のp.183に次のように書いてある。

> The absence of significance relatively to the size of the samples is too often interpreted by the casual reader as a denial of all differentiation, and this may be disastrous.

「…**統計的有意性と社会的重要性との区別は、すべての研究者にとって明らかであるべきだ**…我々には、真の違いが存在するかどうかを判断し、その社会的重要性と**そのコスト**を示す責任が課せられている。**統計的に有意な差が見つからなかった場合、我々は直ちに真の違いは存在しないと結論付けることは正当化されない**。」

— P. 118 of JW Tyler, Educational Research Bulletin, Mar. 4, **1931**

「自動化された意思決定の最も悪質な乱用の一つは、観察された差に対する非有意なP値に基づいて、臨床治療が同等であると断言されるときに起こる…我々は、形式的な仮説検定のように、我々の決定を自動化しようとするいかなる試みにも抵抗し続けなければならない。」

— Claire Weinberg, “**It’s Time to Rehabilitate the P-value**”, Epidemiology 2001; 12: 288-290.

Brown et al.は、HR 1.7 [1.1, 2.6]を持つ**4つ**の先行コホートの自身のメタ分析で、同じリスク増加を示す**HR 1.7 [1.1, 2.6] を報告していたが…**

- **彼らは、自分たちの新しい研究をそれらの研究と統合しようとはしなかった。**
- **そして彼らは、HR 1.74 [1.19, 2.54]を持つ16のコホート研究と、HR 1.95 [1.63, 2.34]を持つ5つの症例対照研究からなるHealy et al.による2016年のメタ分析を引用しなかった。**

なぜ、曝露者における60-70%高いリスクという一貫した関連についての議論がないのか？

それは、ほとんどの人がこの高度に再現された関連が純粋な交絡であると確信していたからである:

- Medscape 2017: 「**妊娠前および妊娠中の抗うつ薬の使用は、自閉症やADHDを引き起こさないことが新しい研究で示された。3つの研究は、妊婦における抗うつ薬の使用が子供の自閉症スペクトラム障害(ASDs)の原因である可能性は低く、以前の研究で見られた関連は交絡因子によるものであった可能性が高いことを実証(demonstrate)している。**」

支配的な社会的バイアスは、あたかもすべてのインセンティブが効果を論破するのではなく「発見」することにあるかのように語る。このメタバイアスは、「再現性の危機」に関する文献に蔓延しており、トピックや著者によるインセンティブの違いを無批判に無視している。

- Brown et al.の例は、CI(信頼区間)が最終的に1を含むまで調整を続けることで**幅を広げようとするCIハッキング**の様相を呈している(たとえ初期のCoxモデルを超える調整が、バイアスを除去せずに分散を膨らませる過剰調整の様相を呈していても)。

ここでのポイントは、出生前のSSRIがASDを引き起こすと主張すること**ではない**(巨大なトピックだ！)。むしろ、

- **「スピン」が「分岐する小道の庭」を通る駆動力であるということだ: 「客観的な」統計は、好ましい因果ストーリーに基づいて知覚され、選択され、報告され、そして利害の大きい状況では、政治的および訴訟上の懸念に基づいて報告される。**
- **例は健康科学および医学科学の至る所にある** — これはあなたを怖がらせるはずだ！
- そうでないと装う統計学の訓練は、この操作を覆い隠し、助長する。

訳註: “分岐する小道の庭”：この語句はホルヘ・ルイス・ボルヘスの1941年の短編小説 The Garden of Forking Paths (https://en.wikipedia.org/wiki/The_Garden_of_Forking_Paths) のタイトルに由来し、「無数の可能性が分岐し続ける状況」や「過去・未来が何通りにも展開しうる世界」を象徴している。

- **「我々」(研究者、査読者、編集者)が信じたいと願う因果ストーリーが、分析の選択と出力の解釈に因果的に影響を与える。その結果、報告書はしばしばそのストーリーを正当化するためのこじつけ(lawywering for)として機能している。**
- **この問題に対する盲目の主な源泉は、統計学や「メタ研究」の専門家たちが、自分自身の認知バイアスや政治的バイアス、訓練の不備、さらには統計の開発者、教員、ユーザー、消費者の不備を無視していることにある。**

- **ロマン主義的・英雄的ファンタジーとしての科学: 社会的な結果にかかわらず、事実の発見と正しい事実の普及に献身する…**
- **しかし、結果をいっさい顧みずにすべての正しい事実を普及させる者は、ほとんど存在しない。**
- **厳しい現実: 統計学の多くは、主要な社会ネットワークの利害に奉仕しており、事実の描写を歪めてプロパガンダに変え、そのネットワークの価値観や特殊利益に従って社会を方向づける役割を果たしている。**

例: 専門家による「EBM」が、ランダム化比較試験(RCT)を「ゴールドスタンダード」として延々と宣伝しているが、RCTは以下の理由でそのようなものではない:

- 倫理的および法的責任上の理由から高リスク患者が除外されること、そして実際の副作用を持つプラセボ製剤による、**巨大な一般化バイアス**。
- **有害作用を見極めるには数が少なすぎ、追跡期間が短すぎるため、非有意性が「再現性の失敗」として報告される。**
- **隠されたプロトコル違反に加え、選択的な出版、報告、議論…**

訳註: EBM は evidence-based medicine の略で「根拠に基づく医療」を意味している。

1 June 2022 Greenland – Reforming Statistics 30

典型的な例: Vallejos et al.によるRCT 'Ivermectin to prevent hospitalizations in patients with COVID-19' BMC ID 2 July 2021…

- 要旨: OR = **0.65**; 95% CI **0.32, 1.31**; p = .23 「イベルメクチンは入院予防に有意な効果はなかった」と報告された。
- Gideon M-K "Health Nerd" (Medium 16 July 2021) は、この試験が「死亡に対するイベルメクチンの利益を見出さなかった」と書いた — **しかし**、Vallejos et al.の5ページ目には: OR = 1.34, 95% CI **0.30, 6.07** (**イベルメクチン群で4人、プラセボ群で3人の死亡**)。
- **この試験は、何かを示すには規模が小さすぎた！**

1 June 2022 Greenland – Reforming Statistics 31

MedPage Today 2021年5月21日の調査: 「IBD試験におけるベイト・アンド・スイッチ？主要評価項目がしばしば未報告または途中で変更される」

- 「公表された結果を持つ57の第III相試験の分析によれば、**その半数[~50%]は、事前に規定された主要評価項目のうち少なくとも1つを報告しなかった(17.5%)か、または少なくとも1つが説明なしに変更されていた(33.3%)**。」

他の研究では、多くの登録済み試験が、出版する意向を表明しているにもかかわらず、決して出版されないことが判明している。

1 June 2022 Greenland – Reforming Statistics 32

実証的事実: **我々は皆、愚かである**(もしも腐敗していないならば)

エイモス・トヴェルスキー: 「**私の同僚は人工知能を研究しているが、私は自然な愚かさを研究している。**」
「**ほとんどの素人が陥る単純な誤りがあるときはいつでも、専門家が陥る、同じ問題のわずかに洗練されたバージョンが常に存在する。**」

例: P値 = 「帰無仮説が成立する確率」 vs. P値 = 「偶然だけで関連が生じる確率」— しかし「偶然だけで」**は帰無仮説の成立を意味している**！

訳註: エイモス・トベルスキー (https://ja.wikipedia.org/wiki/%E3%82%A8%E3%82%A4%E3%83%A2%E3%82%B9%E (Amos Tversky (https://en.wikipedia.org/wiki/Amos_Tversky)) はイスラエル出身の心理学者。スタンフォード大学教授。ダニエル・カーネマンとともに行動経済学を作った。

1 June 2022 Greenland – Reforming Statistics 33

実証的事実:
高名な人達の無能さは普通のこと(the norm)だ。

トヴェルスキー: 「何かを知らないかもしれないと考えるのは恐ろしいことだが、概して、世界が、**自分たちは何が起こっているかを正確に知っているという信念を持つ人々**によって動かされていると考えるのは、もっと恐ろしいことだ。」
— これは研究**および**方法論においても同様に真実である！

- **Covid-19パンデミックは、我々に鮮明な現実世界の例を供給してくれた — しかし、それらの例が誰であるかについての合意はない。**

- カーネマン: 「**人々は、正当化されるよりもはるかに高い確率を、自分たちの意見の真実性に割り当てる。**」
- **純粋な意見を神聖化することにより、ベイズ的手法は、事前分布のスパイクや「専門家知識に基づく事前分布(elicited priors)」(バイアス、文献の誤読、個人的偏見の要約表現)を介して、統計学にさらなる乱用の余地を与えてしまう。**
- 例: **Pr(帰無仮説)=0.5は「無差別」だと主張することは、無差別ではなく、巨大なゼロバイアスである。**

訳註: 事前分布のスパイク(prior spike)は効果を表すパラメータ θ に関するゼロ仮説 $\theta = 0$ に台を持つ事前分布の形状を表している. そのような事前分布の密度関数はデルタ関数 $\delta(\theta)$ になる.

さらにもっとカーネマンより:

- 「我々は明白なことに対して盲目になり得るし、我々はその盲目さに対しても盲目である。」

そして、ソフトサイエンスにおける統計学に最も関連すること:

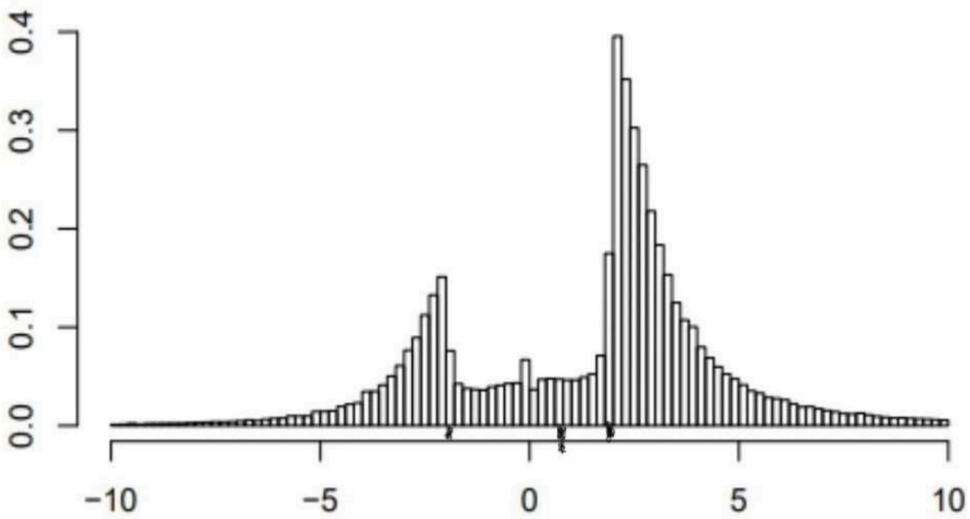
- 「…**妥当性とスキルの錯覚は、強力な専門的文化によって支えられている。我々は、人々が、たとえそれがどんなに馬鹿げた命題であっても、同じ考えを持つ信者のコミュニティによって支えられているとき、揺るぎない信念を維持できることを知っている。**」

— 参照: 有意性検定のあらゆる擁護…

例: 「**もしも効果に関するp値が雑誌のp値の閾値よりも大きいならば、編集者は直ちにその論文を却下することができる。**それによって雑誌は(説得力のない)論文にこれ以上時間を費やすことから救われる…もしも結果が統計的に有意であれば、これは、研究された効果について、その効果を報告する論文を**出版のために検討する価値があるほどの**証拠の重みがあるということに他ならない。」

— これは1920年代のFisher？
いいえ、これは2021年の統計学です:
McNaughton, The War on Statistical Significance.

1950年代に指摘された事実を無視している: 研究結果に基づく**あらゆる**選択的報告は、入手可能な結果の分布を全体の分布に対して歪めるだろう。



推論の基礎をカバーするとされるいかなる指導も、
以下のような社会的妄想やバイアスに対処するために
認知科学を含む必要がある:

- **ゼロ主義(nullism)**: 簡潔性への我々の要求(ゼロへの収縮)を現実と混同すること。
- **二分法への執着(dichotomania)**: 要約(単純化)と決定に関する我々の要求を白黒つけたがる我々の傾向と混同すること。
- **現実との混同(reification)**: 推論・推測・意思決定に関する**形式的な方法**だけで、現実世界における推論・推測・意思決定を十分に行えるという信仰。

ゼロ主義(nullism)は、物理崇拝者の間での
二セ懐疑主義(ゼロ仮説への実証的に擁護不可能な確信)
として長く輝かしい歴史を持っている：

- 「**空気より重い飛行機械は不可能である**」
  — ケルビン卿 1895年、1902年に繰り返される
- 「**大陸移動説は論外である**」
  なぜならば十分な強さを持つメカニズムが存在しないからだ
  — **ハロルド・ジェフリーズ卿**、**スパイク事前分布(＝形式化された過信)**の創始者である地球物理学者。
- 膨大な証拠があったにもかかわらず、タバコが肺癌を引き起こすことに反対し続けたフィッシャーも参照せよ。

訳註: "pseudo-skepticism"を「擬似懐疑主義」と直訳せずに、より否定的に聞こえる「二セ懐疑主義」と訳した。意味的には「健全な科学的懐疑主義とは異なる不健全な懐疑主義」というような意味だと思われる。

- **ゼロ主義**への反対: 現実は、単純または決定的である義務を負っていない。
- **二分法への執着**への反対: 重要とされる意思決定の多くは二者択一ではなく、二者択一である**べきではない**: オーブンの設定温度は？サーモスタットの設定温度は？**薬の投与量は？**
- **隠れた現実との混同**: 研究者は、広範なモデルの不確実性を無視した「推論」を日常的に発表している — 彼らは、自分たちのモデルにおけるすべての単純化を無視する根拠を知らず、それらについて考えようともしない。

他の多くの認知バイアスが、
設計、分析、報告、出版バイアスに寄与している
https://en.wikipedia.org/wiki/List_of_cognitive_biases

以下のすべておよびそれ以上が、推論を節度あるものにするための基本訓練に含まれるべきである:

- **固着バイアス(anchoring)**: 修正された後であっても、見かけ上の合意や望ましいが誤った信念に固執してしまうこと。
- **確証バイアス(confirmation bias)**: 望ましい証拠に選択的に注目し、望ましくない証拠を無視する傾向。
- **礼節バイアス(courtesy bias)**: 相手を不快にさせる批判については、あいまいに表現してしまう傾向。

- **代替案を検証しないこと** (「適合性バイアス(congruence bias)」)
- 好ましくない証拠への**選択的批判 (selective criticism)**
- 選択的な仮定・説明・データを通じて望む結論に至る**選択的推論 (selective reasoning)**
- **ダニング＝クルーガー効果 (Dunning–Kruger effects)**: 専門知識が低いほど、自分の能力を過大評価する傾向(例：研究者が自分の統計的専門知識を過大評価する場合、医療雑誌の統計編集者など)
- **過信、妥当性の錯覚 (overconfidence, validity illusions)**: 使用する数学(思考実験)上で正確な方法や判断を、現実世界でも同じくらい正確だと考えてしまう傾向

- **慣れ親しみバイアス (familiarity bias)**: 慣れ親しんだ方法への過度に依存、代替的な手法の無視 (「これで研究費も論文も得られるのだから、変える必要はない」)。
- **縄張り(排他的)バイアス (territorial (exclusionary) bias)**: 慣れ親しんだ方法を唯一正しいアプローチとして推し進め、自分の権威を守り、競争相手の台頭を防ぐこと。
  「ストリクトリー・ボールルーム効果」: 十分に学んで使い込んでいないことについては権威になれない。
- **集団思考と群集行動バイアス、例えば反復バイアス (groupthink and herd-behavior biases such as repetition bias)**: エコーチャンバー効果。集団内での相互強化によって証拠が実際より沢山あるかのように見えてしまうこと。

訳註:

- 「ストリクトリー・ボールルーム」(“Strictly Ballroom”)は1992年のオーストラリア映画 (邦題『ダンシング・ヒーロー』、ダンシング・ヒーロー (映画) - Wikipedia (https://ja.wikipedia.org/wiki/%E3%83%80%E3%83%B3%E3%82%B7%E3%83%B3
  オーストラリアの競技ダンス界を舞台に、型にはまったダンスのルールや慣習に反発し、自分のスタイルで踊ろうとする若いダンサーのストーリー。保守的で伝統に固執するダンス界の「型破り」を描く。
- 「反復バイアス(repetition bias)」は「何度も繰り返し引用・使用される仮説や結果が、検証の弱さにかかわらず妥当だとみなされ易くなること」という意味。

- **心の投影の誤謬(mind-projection fallacies)**: 本来意味を持たない数量に、態度・意見・価値・推論・判定・意思決定を浸透させること。
  — 「有意性(significance)」「信頼性(confidence)」「厳密性(severity)」のような**価値識別子**を言葉の本来の意味を表せない狭義の数学的概念に用いることで、この誤謬は統計学的議論において蔓延している。
- ナンセンスの最たる例：「**P値は証拠を過大評価している。**」 P値はモデルから導出された参照分布(例えばカイ二乗分布)における検定統計量の値の位置を表しているだけである。**証拠の過大評価はどれもそれを見る側によるものである。**

これらはは絶対的または明確な類型群(categories)ではなくむしろ、同僚(どんなに善意であっても)や「専門家」や何よりも**自分自身**に油断させられたり騙されたりしないための経験則的な警告群(heuristic triggers)である。例えば、

- ダニング＝クルーガー効果という**医療評論家の間で蔓延している過信バイアスの一形態** (統計的手法についてコメントするときだけとは限らない): 我々は自分の専門分野については非常によく理解しているかもしれないが、その専門知識をすぐに他の分野に一般化できないことには気づいていない。これは、自分の専門分野に近いと思っている話題であっても、実際には自分が認識しているよりもはるかに多くの文献が存在する場合に当てはまる。

「出版された研究結果の大半が誤りである」ことの主な理由はシステム的な問題にある:

- 誰もがそうであるように、**統計学の教員達 (stat instructors)**、ユーザー達、消費者達も**二分法への執着 (dichitomania)**、**ゼロ主義 (nullism)**、**現実との混同 (reification)**という病気に罹っている: **彼らはゼロ仮説に関する真か偽かの結論を切望し、そのため過度に単純化されたモデルからそういう結論を受け入れてしまう。**
- しかし、「ソフトサイエンス」への応用において、**観察は(RCTの場合でさえ)決してそのような絶対的な確実性を提供することはできず、不確実性の正確な評価さえ提供できない。**

- 統計学は、洗練された**意思決定論**を提供することで我々の渇望に応え、ユーザーには観察が決定的なリスクと不確実性の評価を提供できる**かのように見せかける**。
- **「信頼区間」は、応用場面におけるすべての不確実性の原因を捉えているかのように見せることによって、これらの錯覚を永続化している。しかし「信頼区間」が捉える不確実性は「信頼区間」の計算に使われたモデルを前提としたものだけである。**
- **さらに悪いことに、標準的なプレゼンテーションでは、モデルの不確実性の無視についてほとんど言及されない！**

1 June 2022 Greenland – Reforming Statistics 48

- 統計学は、**変動やバイアスの形態や原因**に関する深い不確実性に正面から向き合わずに、すべての不確実性を偶然のゲーム(**既知の**形の確率分布からの無作為抽出)から来たかのように扱う**遊戯的誤謬 (ludic fallacy)**にもあっさり陥っている。
- **これらの問題は、因果的思考の誤りと認知バイアスに関する教育・理解が、データからの健全な科学的推論を促進すると主張するあらゆる専門分野において不可欠であることを強調している。**

1 June 2022 Greenland – Reforming Statistics 49

- **数学化はこれらの誤謬における過信を増幅し、統計理論を現実世界に関する誤情報の源泉(fountain)にしてしまう**:
- **数学的導出は、その前提(例えば少数のモデル候補群が現実をうまく近似できるという仮定)が成り立っている場合に限って、「最適性」などの結論に関する確実性を保証するにすぎない。**
- それにもかかわらず、数学的導出の結論はあたかも自明の真理であるかのように扱われ、従来の訓練・教育・実践に関する思い込みによってその感覚は強化される。**その結果生じる共有された認知バイアスは、社会的なフィードバックループによってさらに強化されていく。**

1 June 2022 Greenland – Reforming Statistics 50

- **その結果は、技術的に最も熟練した支持者者達の議論における集団思考、隠れたバイアス、循環論法である！**
- 典型例はネイマン流およびベイズ流の優位性を主張する一般的な議論(ネイマンやベイズ自身の著作に見られるものよりも悪い)。
- 例: 現実世界における結論や意思決定において、校正(calibration)やベイズ的一貫性(Bayesian coherency)を最優先の指針として要求すること。これらは**それぞれの表現体系内における**指針にすぎず、**取り込まれていない文脈**を無視することで破局を招く可能性がある。

1 June 2022 Greenland – Reforming Statistics 51

数学的正当化を十分な実践的正当化であるかのように扱うことをやめよ

- どんなに複雑に見えようとも、数学の結果は、現実の実践よりもはるかに単純な理想化された設定で手法をテストするための、単なる思考実験にすぎない。
- これらの単純なケースでの性能は、問題についての警告や手法の改善提案とともに、実践のための指針を提供することができる。**しかし、**
- 単純な設定で見られる問題は、複雑な設定では悪化する可能性があり、
- 数学的な「最適性」の結果も、数学的な設定で問題が見つからないことも、実際の応用で良好な性能を保証するものではない。

1 June 2022 Greenland – Reforming Statistics 52

価値バイアス(value bias)は統計的方法論に浸透しており、
最も頻繁には**ゼロ主義(nullism)の形で現れる**
(ゼロ仮説を「受容する」ことに偏った価値バイアス):

- ある方法論がすべての利害関係者が共有していない費用/便益の前提を組み込んでいるとき、その方法論が**価値バイアスされている(value-biased)**という。
- **これらのバイアスは通常統計的伝統や数学への固執によって公には認識されにくくなっており、NHST(帰無仮説有意性検定)やそのベイズ的類似における不明瞭な**

費用/便益の前提に潜む価値観を覆い隠している。

訳註: 「価値バイアス(value bias)」は「分析や解釈に、研究者や社会の価値判断(何が重要か、望ましいか)が入り込み、結論を歪めること」という意味。

1 June 2022 Greenland – Reforming Statistics 53

- 例: **ゼロ仮説を一貫して検定仮説(test hypothesis)として用い、ゼロ仮説達と検定仮説達を区別できなくなること(フィッシャーに起因する誤り)。**
- これは**ゼロ主義**の一例であるゼロ仮説を支持する方向への価値バイアスであり、次のような人達を有利にしてしまう: ゼロである方が都合が良い人達(製品監視などに見られる)、ゼロであることに形而上学的信念を持つ人達(簡約化のヒューリスティック(parsimony heuristics)と自然法則を混同するニセ懐疑論者(pseudo-skeptics))。
- 多くの研究者は、**任意の効果サイズに対してP値("tested")を付与できる**ことに気づいていない。

訳註: 「P値("tested")」はおそらく「観察データの値から計算されたP値」を意味するものと思われる。検定された仮説(tested hypothesis)に関するP値は観察データの値から計算されている。モデルの確率分布に従って生成されたモデル内確率変数としてのデータから計算されたモデル内確率変数としてのP値とは異なり、観察データの値から計算されたP値は定数になる。

1 June 2022 Greenland – Reforming Statistics 54

- NHSTを通じて、ゼロ主義(nullism)はネイマン＝ピアソン検定の不可欠な部分として教えられてきた — たとえそれがそうではないとしても！

Neyman, Synthese 1977の104頁, 106頁より(強調は筆者による):

- 引用「**状況や研究者達の主観的態度に応じて、ある誤りは他の誤りよりも避けることが重要に見えることがある。避けることがより重要な誤りを「第一種の誤り(error of the first kind)」と呼ぶ。**」(「タイプI」エラー、アルファエラー、**検定仮説 H (test hypothesis H)の誤った棄却**)

1 June 2022 Greenland – Reforming Statistics 55

- 引用「それの不当な棄却が第一種の誤りを構成する[**仮説**]は、**『検定仮説 (the hypothesis tested)』と呼ばれるだろう。**」
  — **この記述が、検定仮説 H (test hypothesis H)として非ゼロ仮説をどのように許容しているかに注意せよ。**
- 引用「化学物質Aの製造者の立場からすると、Aが発がん性を持つと主張する誤りは、Aが無害であると主張する誤りよりも(あるいはより一層)避けることが重要である場合がある。このことより、**製造者にとっての『検定仮説 (hypothesis tested)』はしばしば『Aは発がん性を持たない』になる。**」

1 June 2022 Greenland – Reforming Statistics 56

- 引用「**一方、化学物質Aの将来の使用者にとって、検定された仮説(hypothesis tested)は明確に『Aは発がん性を持つ』になる。**実際、この使用者は、この仮説を棄却する際の誤りの確率が非常に小さくなることを望むであろう。」
  — **これは、統計的検定を教えたり、推奨したり、使用したりする者は誰であっても、使用するカットオフ(P値、ベイズファクター、尤度比などいずれであっても)だけではなく、検定仮説 H (test hypothesis H)の選択を正当化する必要があることを意味する。**

1 June 2022 Greenland – Reforming Statistics 57

- このようにネイマンは、検定仮説(test hypothesis)の選択における価値観の役割と、それが利害関係者間でどのように(そしてしばしばどのようにでも)変わり得るかについて明確な説明を提供した。
- **しかし、多くの「オピニオンリーダー達」は、ゼロ主義に基づく検定のみを行うという硬直した慣行を維持している。**これは、粗雑に単純化された生物学モデルへの信頼、選択的観察からの一般化、簡約化のヒューリスティックをあたかも形

而上学的原理であるかのように扱うこと、そしてもちろん**スポンサーやイデオロギーへの奉仕などの隠れた価値評価**に基づいている。

- 我々は、存在するいかなる効果も「十分に小さい」ため、それを無視するコストが許容できると確信しているかもしれない **— しかし、それこそが価値判断だ！**
- 統計的検定は、治療の優越性、劣等性、非優越性、非劣等性、さらには同等性のために構築することができる — しかし、そのような宣言を定義する効果サイズを人為的に正確に指定する必要がある。
- そのような「境界」効果サイズに関するP値が提示されるべきである — そして**P値グラフ(P-value graph)は、そのようなすべての選択肢に関するP値を示すことができる。**

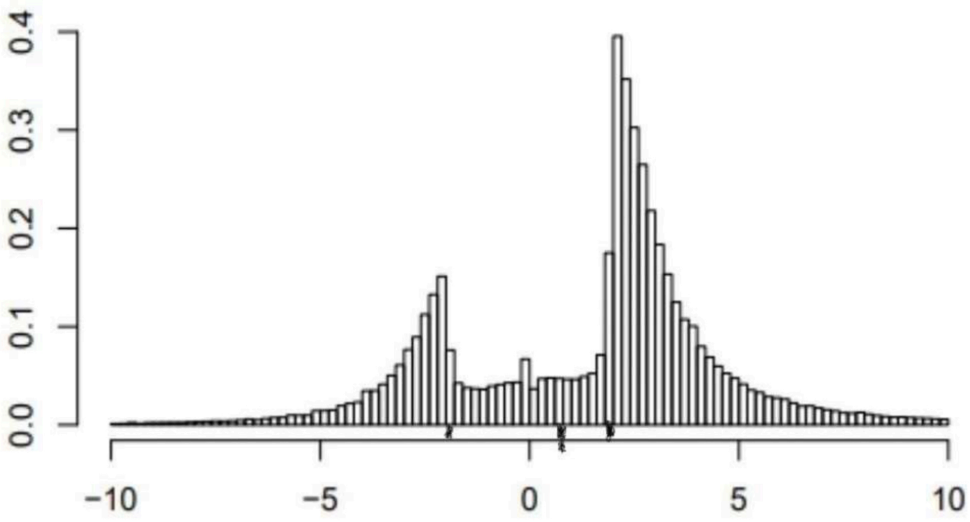
訳註: データの値から決まる仮想的効果サイズにP値を対応させる関数(P値関数、P-value function)のグラフをP値グラフ(P-value graph)と呼ぶ。

多重性調整は価値バイアスを悪化させる

- 多重性調整では伝統的に、**ゼロ仮説達が成立していること(joint ensemble null)**を誤って棄却してはならない最も重要な仮説とみなし、**ゼロ仮説達全体が成立していること(the entire ensemble of nulls)**に0.05の第一種過誤率を適用する。
- **このように多重性調整では、偽陽性のコストは常に偽陰性のコストより大きく、そのコスト比は仮説の数が増えるにつれて常に増加する、と仮定している — この価値評価は、有害事象を監視する製薬会社には当てはまるが、患者には当てはまらない！**

- **ゼロバイアスはベイズ統計の文献の多くにも見られ**、そこでは**ゼロスパイク(null spikes)**がパラメータがゼロとの「違いは無視できるほど小さい」という信念を誤って表現するために用いられている。
- 多くの医学研究の場面では、ゼロの周りに事前確率を集中させることには、実際のデータに基づく根拠がない。実際、**事前スパイク(prior spikes)は通常本来の事前情報と矛盾する。**例えば、**潜在的な医薬品は、まさにそれらが標的とする生理学的システムに影響を与えることが知られているからこそ研究される。**

例: 再び、van Zwet & Cator 2021の図1。一部のベイズ主義者は、補完された曲線が右に歪んでおり、75%以上が0を上回っているにもかかわらず、推定値を0に向かって縮小させるだろう。経験ベイズ主義者は、代わりに、推定されたトピック固有の平均への縮小を使用するだろう。



第I部まとめ:

- **数学的枠組みの盲目的な受容、「偉人たち (great men)」とその概念的誤りの神格化、そして認知的問題の無視が、統計学の訓練や実践およびその結果として公的情報の蓄積の中核を腐敗させてきた。**
- **「再現性の危機」ヒステリーはこの問題をさらに悪化させている。その原因は、ゼロ主義(nullism、対立仮説の検定の無視)、無意味な二分法への執着(dichotomania)、有害なモデルと現実の混同(model reification)にあり、これらすべてが帰無仮説有意性検定(NHST)とゼロスパイク付き事前分布(null-spiked priors)に制度化(enshrined in)されている。**

- 統計的方法論によって推奨されてきた持続的な実践的誤り: **現実世界の推論を、たった一つの研究、たった一組の背景仮定、たった一つの形式的推論体系(formal reasoning system)、たった一つの解釈だけに基づく演繹によって構築すべきだ、という考え方。**
- 著者達の多くは、多様な研究デザイン(単なる「再現実験 (replications)」に限らない)や多様な仮定(感度分析)の必要性を認めてはいるが、**多様な方法論と解釈の必要性については自覚がないように見え、それに反対する者さえいる。**

訳註: 形式的推論体系(formal reasoning system)は頻度論的もしくはベイズ的な特定の流派の推論体系を意味するものと思われる。

- **統計的ルール達は悪い実践をさらに悪化させる可能性がある。**なぜならば、ルール達の理論は、我々が厳密に管理された実験の完全な解釈だけを用い、誤りのコストを明確に把握していることを前提としているからである。
- しかし**ソフトサイエンス研究における多くの「データ分析」は、統計的出力に対して意思決定規則を適用することに終始している。その際、デフォルト設定に潜む価値判断を含む性質(value-laden nature)が利用者や読者のほとんどに認識されていない。**例えば、P<0.005 を「関連あり」の報告のための要件としたり、P>0.05 を「関連なし」と誤解したりする。

むき出しの心理社会的事実:

- **統計的出力の「客観的」な記述の大半は実際には主観的な解釈であり**、通常は意思決定規則があたかも推論規則であるかのように誤って提示されている。それらは実際には推論規則ではない。なぜならば特に意思決定規則は正当化されたコスト関数を必要とするからである。
- **さらに悪いことに、ほとんどの入門書、チュートリアル、教科書に見られる言葉による定義や記述はまるっきり間違っている(flat-out wrong)。** 例: Cassidy et al., “Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly”, AMPPS 2019

改革に向けた、緊急かつ見過ごされがちだが、容易な一歩:

- **統計学が提供できるのはデータ変換だけである**と教える。例: P値、事後確率、区間推定。
- **観察者は、物理的な研究の現実 (physical research reality)の因果モデルを介して、統計に意味を与える。**

これより、正当化された「統計的推論」には以下が必要となる:

- **統計的仮定が物理的な研究の現実 (physical research reality)からどのように導出されるかを示すこと。**
- **データがそれらの仮定を総合した予測と比較してどこに位置するかを示すこと ー まさにそれがP値の役割である！** そこから導かれることは…

改革の一歩:
帰無仮説有意性検定(NHST)の糞の山(dungheap)から
P値を救い出せ

- 有意性検定の批判者も擁護者も、P値を NHST における伝統的な使い方と同一視して誤認している。これはすべてのナイフを「武器」と呼ぶようなものである — **それは道具とその誤用の混同である**(NHSTにおけるP値の使用が思考を殺すのに使われている場合がそれにあたる)。
- その混同は破壊的である。なぜなら**P値はNHSTから切り離して多様な目的に利用できるからである。P値は、モデルの適合度の尺度として利用でき、推定装置としても利用でき、頻度論とベイズ統計学を結びつける論理的な橋を築くためにも利用できる。**

1 June 2022 Greenland – Reforming Statistics 68

二分法への執着と「有意性」を追放したあとに残る
P値の再評価に立ちはだかる課題

- 教員とユーザーはP値が検定仮説H (通常は関連なしまたは効果なしのゼロ仮説)が成立する確率であることを望む。
- P値は通常その確率に近くない。
- **それにもかかわらず、教育や研究の文献は、P値をあたかも仮説が成立する確率であるかのように扱う微妙に誤った言い回し(「P値反転 (P-inversion)」)を助長している。**

1 June 2022 Greenland – Reforming Statistics 69

醜い事実:
「推論統計学」の妥当な解釈は
ほとんどの情報源の手に余るように見える

- **文献は、頻度論的解釈とベイズ的解釈を混同した、P値の不適切な記述で満ち溢れている。**
- **例:「P値は結果が偶然による確率である」のような反転、「P値は偶然の発見の確率である」のような無意味な記述。**
- **信頼区間の多くの記述は、実際には事後区間を定義している。しかも…**
- **95%「信頼」区間は、通常、5%水準の検定以上のものとして扱われることはない。**

1 June 2022 Greenland – Reforming Statistics 70

反転の誤謬には、PP値を「ランダム性」や「偶然だけ」で関連が生じた確率だと誤解することが含まれる。例えば、Harris & Taylor, Medical Statistics Made Easy* (第2版、2008年、p.24–25)では、P値は「**観察された差が偶然に生じた確率**」(偶然だけで？)であると説明している。

- **もしも検定される(「ゼロ (null)」)モデル(効果なし、バイアスなし、モデル化の誤りなし)が正しいならば、ゼロでない差が「偶然だけ」で生じる確率は何か？**
  答え: **100%**

*(「Made Easy (簡単にする)」は「Made Wrong (間違いにする)」の婉曲表現か？)

1 June 2022 Greenland – Reforming Statistics 71

- 健全な分析は結果を非常に曖昧でしばしば非対称なものとして見る必要がある。しかし、
- 証拠と不確実性の概念は、データが関連するとされる明示的なモデルとの関連においてのみ定量化できる。例:
  - モデルの予測とデータの対比 (相性チェック = 「適合度の検定」、**頻度論的**診断におけるもの)、もしくは
  - モデルとデータの統合による予測や賭けの更新 (**ベイズ的**事後確率計算におけるもの)。

1 June 2022 Greenland – Reforming Statistics 72

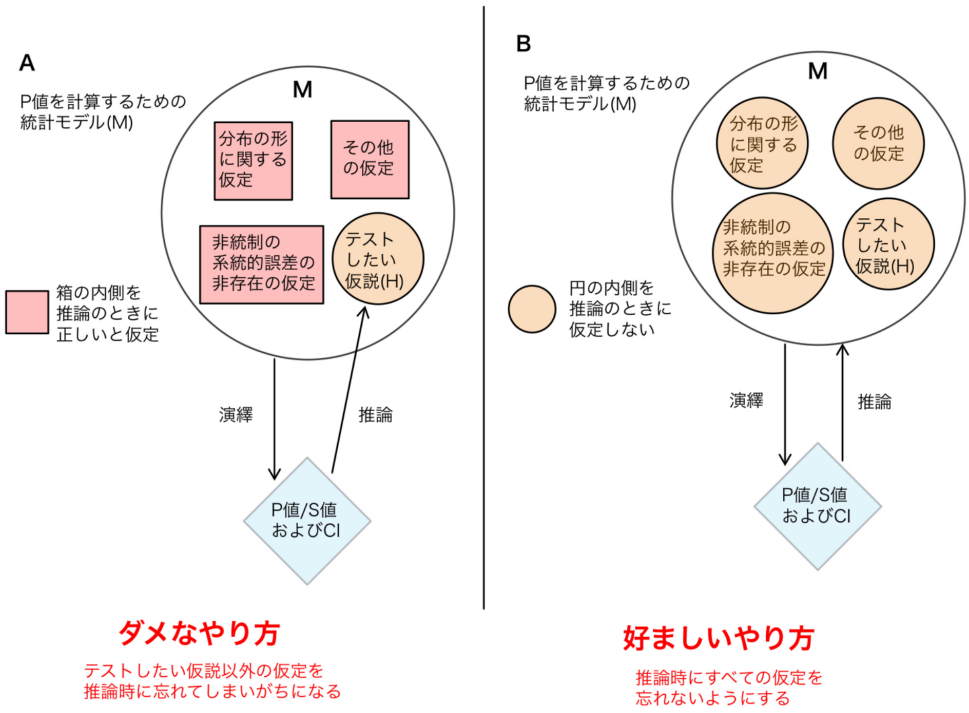
統計学の訓練の再構築:
統計学の「偉人たち (great men)」の過ちと、
彼らが示し、生み出し、助長した認知バイアスを
永続させることをやめよ

- 統計学の教科書は、教員や学生が論理と意味論を十分に理解しており、悪い専門用語(bad terminology)を見抜き、数学的意味と文脈的意味を区別できることを前提としている。
- フィッシャーの全盛期以前からの不満が示すように、**それは決して真実ではなく、20世紀中頃の研究の爆発的増加で悪化するだけだった。**

条件付きの「仮説検定」的解釈から、
条件付きでない記述的解釈に重点を移せ

- 一般的な解釈: 「**P値は、検定仮説Hが正しい場合に、検定統計量が観測値以上に極端な値になる確率である。**」**これは背景の仮定(モデル)を暗黙のままにする。**
- その代わりに、背景の仮定を前面に出す。例: P値 p は検定モデルの下で検定統計量がどのパーセンタイルに位置するかを表す。
- 検定モデルには検定仮説HとP値 p を計算するために使用された他のすべての仮定が含まれる。

Greenland & Rafi “Semantic and cognitive tools to aid statistical science” http://arxiv.org/abs/1909.08583 (http://arxiv.org/abs/1909.08583) より



ダメなやり方

テストしたい仮説以外の仮定を
推論時に忘れてしまいがちになる

好ましいやり方

推論時にすべての仮定を
忘れないようにする

訳注: 上の図は

- Sander Greenland and Zad Rafi, To Aid Scientific Inference, Emphasize Unconditional Descriptions of Statistics, http://arxiv.org/abs/1909.08583v5 (http://arxiv.org/abs/1909.08583v5)

の図1の翻訳版。図1には次の内容のキャプションが付けられている:

図1: P値、S値、相性区間（CIs）の条件付き解釈と条件付きでない解釈
(A) 条件付き解釈：背景モデルの仮定（例：系統誤差が存在しない）が正しいと仮定する。この場合、P値とS値が提供する情報は、検定仮説を対象としている。
(B) 条件付きでない解釈：統計モデルのどの側面も正しいとは仮定しない。この場合、P値とS値が提供する情報は、検定モデル全体を対象としている。

- **検定統計量は検定仮説Hおよび背景仮定を含む検定モデルの予測とデータのずれを測る。**
- **したがって、P値は検定仮説Hだけでなく検定モデル全体を前提にして計算される。**
- **条件付けを外すことは、P値を導出する際に用いられた仮定のどれか1つだけが破れても、そのP値の大きさに影響し得ることを強調する。**P値の大きさに影響するのは**検定仮説H**の破れだけではない。
- **大きなP値は検定仮説Hや検定モデルを支持したり確証したりするものではない**(**証拠の欠如は**モデル違反の**欠如の証拠ではない**)。

訳註: 「(**証拠の欠如は**モデル違反の**欠如の証拠ではない**)」の原文は “(**absence of evidence is not evidence of absence** of model violation)” である。“Absence of evidence is not evidence of absence.” (証拠の欠如は欠如の証拠ではない)という決まり文句がある。この決まり文句はカール・セーガン(Carl Sagan)が地球外知的生命の存在可能性や健全な科学的懐疑主義について語るときに好んで使った。

**誤解を招く伝統的なジャーゴン(スタットスピーク)を打倒し、
統計用語を日常言語と再整合させよ(realign):**

- 「significance (有意性)」(Edgeworth 1885)と「confidence (信頼)」(Neyman 1934)を、P値 p によって測定される「compatibility* (相性もしくは相性の良さ)」に置き換える。P値 p の範囲は 0 (相性の良さが皆無)から 1 (データとP値 p を計算するために使用された検定モデルの相性の良さが完璧)の間になり、検定統計量によって測定される方向で評価される。

*「consistency (整合性)」とほぼ同じ意味だが、他の多くの概念に使われすぎている。

訳註:

- 「スタットスピーク(Statspeak)」は「使い続けると統計学に対する批判的思考が阻害される言語」というような意味だと思われる。これはジョージ・オーウェルの小説『1984』に出てくる「ニュースピーク(Newspeak)」に似ている。ニュースピークは全体主義体制で国民の思考を統制するために作られた架空の言語の固有名である。ニュースピークを使い続けると全体主義体制に対する批判的思考が阻害される。
- 統計学には「consistent estimator (一致推定量)」という専門用語があるので、安易に consistency という単語を使い難い。

- **それはなぜか？ その理由は典型的な現代の統計学ユーザーは言葉に依存しているからだ ― 彼らにとって数学は、資金を得て出版するために信仰しなければならない単なる象徴的な呪文にすぎない。**
- **「それは単なる意味論の問題だ」という態度は、意味論によって伝えられる本質的な類推的情報を無責任にも把握できていない。**この失敗は、数学的能力のある人達の間でよく見られ、彼らは構文と演繹を類推プロセスよりも上に置くか、現実と数学との間のマッピングにおける類推の役割を軽視または見過ごす。

**任意の検定仮説(test hypothesis)を
「帰無仮説 (null hypothesis)」と呼ぶ**

フィッシャーの誤りを繰り返すな
(この誤りはゼロ方向バイアス(nullistic bias)を公然と招く)

英語の辞書における「Null」の意味:

- Oxford: 形容詞 2. **値ゼロを持つ、ゼロに関する**; 名詞 1. **ゼロ**
- Merriam-Webster: 形容詞 6. **ゼロの、ゼロである、ゼロに関する**; 名詞 7. **ゼロ**

代わりに、ネイマンの用語である**検定仮説 (tested (or test) hypothesis)を用い、点ゼロ仮説達(point null hypotheses)の代わりに、方向性のある(directional)仮説、ゼロ以外(non-null)の仮説、区間(interval)仮説を検定することを重視せよ。**

ネイマンの「信頼のトリック (confidence trick)」を排除せよ

- **高い「信頼」を割り当てることは、高い確率を割り当てることと区別できない。**
- だから、「信頼区間 (CI)」を**相性区間 (compatibility intervals)に改名・再概念化し、P > 0.03 のような相性の良さの基準の下でデータと比較的相性が良いパラメータの値の範囲を表すと解釈するべきである。** (後で示すように、P > 0.03 という基準は区間内のパラメータの値の否定についてコイントス約5回以下程度の証拠しかないことの要請になっている。)
- **これは計算上および数値上の変更を一切伴わない！ すべては認識の問題である…**

「相性 (compatible)」は「信頼 (confidence)」よりも
ずっと慎重である (論理的にもはるかに弱い):

- 我々のデータと相性が良い可能性(モデル)は常に無数に存在する。**ほとんどは想像もされておらず、現在の知識では想像さえできない。**
- 我々は、ケルビンやジェフリーズのような「偉人たち (great men)」による、後に受け入れられた事実に対する独断的な否定を思い出すべきだ。
- 「信頼 (confidence)」は信念 (belief)を意味し、CIを事後信用区間(credible posterior interval)として扱う反転の誤謬 (inversion fallacies)を助長する。対照的に…

訳註: パラメータθの100(1-α)%事後信用区間(credible posterior interval)は事後分布に従う確率変数としてのパラメータθの値が含まれる確率が100(1-α)%になるような区間である。

相性の良さは信頼の根拠にはならない:

- **偽りのストーリーがデータと相性が良くてかつ効果的な介入につながることがある。**
- 例: 「マラリアは、沼地の周りの地面近くに溜まる悪い空気によって引き起こされる。」
- 示唆される効果的解決策: 高床式の住居にする、湿地を排水する — マラリアと相性が良い原因(悪い空気)と実際の原因(蚊)は、両方ともそれらの介入によって減少する。
- しかし、**そういうストーリーへの信頼(confidence)は最終的に誤解を招くだろう。**例えば、そういう信頼は蚊帳の使用を遠ざけてしまう。

問題：信頼区間(CI)の宣言された(「名目上の(nominal」)
被覆率とは、純粋に仮想的な頻度特性にすぎず、
我々はそこに「信頼」を置くべきではない！

- **「信頼」は、アルゴリズム的に生成された区間が「真の値」を被覆する現実における相対頻度が、生成器(generator)について宣言された通り**(例えば95%)**であることを、我々が確実に知っていることを要求する。**
- **しかし、現実の生成器に関する頻度は未知であるため、そのような信頼は正当化されない。**

- したがって、宣言された被覆率は**仮想的な**データ生成アルゴリズムからの繰り返しの標本抽出に関する性質に過ぎず、我々が確信している因果ストーリーに関する性質**ではない**。

対照的に、相性の良さ(compatibility)は
データとモデルの間に観察された関係にすぎない

- 相性の良さ(compatibility)とは、データセットが検討中のモデルから導出されたデータ生成アルゴリズムから来た場合に、期待される位置から(検定された方向に沿ってパーセンタイルで見て)「それほど離れていない (not far)」ことを意味するにすぎない。
- 95%の相性区間(もしくは相性領域)は、検定された方向で $p > 0.05$ となるすべてのモデルに関する結果を示す。これは、後で説明するように、単純なコイントス実験に翻訳すると「相性がかなり良い(of “high compatibility”)」領域である。

Brown et al.によるJAMA 2017の論文 “Association between serotonergic antidepressant [SSRI] use during pregnancy and autism spectrum disorder [ASD] in children” (妊娠中のセロトニン作動性抗うつ薬[SSRI]使用と児の自閉スペクトラム症[ASD]との関連)の**誠実な報告**は次のようになるだろう:

- 要旨: Coxモデルでの調整済みハザード比(HR)は **1.59** (95% 相性限界 (CL): **1.17, 2.17**)であった。IPTW HDPSを用いると、ハザード比の推定ははるかに精度が低くなり、HR **1.61** (95% CL: **1.00, 2.59**)となった。
- 結論: **HDPSモデルの下では、データは、胎児期の (*in utero*) SSRI曝露と子どものASDとの関連については、等倍から2.6倍の上昇の範囲とかなり相性が良いように見える。**

訳注:

- Brown, Hilary K. et al., Association Between Serotonergic Antidepressant Use During Pregnancy and Autism Spectrum Disorder in Children, JAMA 2017 Apr 18; 317(15):1544-1552. https://doi.org/10.1001/jama.2017.3415 (https://doi.org/10.1001/jama.2017.3415)
- “compatibility limits”を「相性限界」と訳した。相性限界は相性区間の両端の値を意味する。これは信頼区間の両端の値を信頼限界というのと同じ。

Vallejos et al.によるBMC ID誌2021年7月2日の論文「イベルメクチンによるCOVID-19患者の入院予防 (“Ivermectin to prevent hospitalizations in patients with COVID-19”)」の**誠実な報告**:

- 要旨：入院オッズ比は **0.65** (95% 相性限界 (CL): **0.32, 1.31**)で、死亡オッズ比は **1.34** (95% CL: **0.30, 6.07**)であった。
- 結論：**結果の精度が低過ぎるので効果の大きさや方向を決定するには不十分であった。**データは、イベルメクチン群の入院オッズについてはプラセボ群と比べて68%低下から31%上昇までの範囲と、死亡オッズについては70%低下から500%上昇までの範囲とかなり相性が良いように見えた。

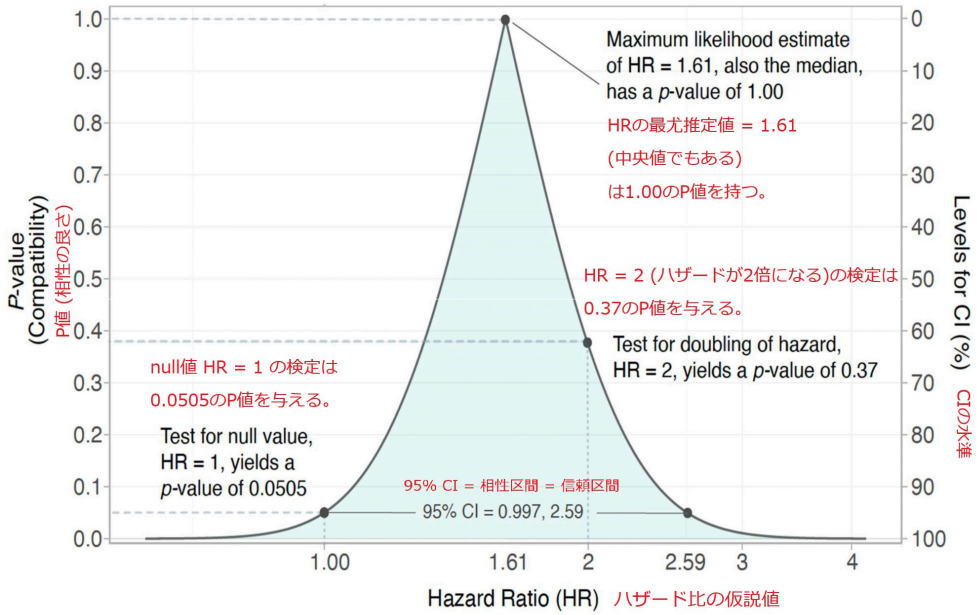
P値を推定ツールとして扱わない
という巨大な誤りを繰り返すことをやめよ
(ゼロ方向バイアス(nullistic bias)を公然と招くもう一つの誤り)

「有意性検定[意味: P値]**と推定の区別は人為的である」**
– エドウィン・ジェインズ、ベイズ的情報理論家

- 実際、**この区別は、検定や意思決定を仮説やモデル達のスペクトル全体の中のたった一つの仮説(ゼロ仮説(the null))とモデルのみに集中宇させるという点で、完全に破壊的であった。**

- **P値と相性限界(CL)をPグラフ全体(entire P-graph)**(P値関数)**上の点を指し示すものとして視覚化せよ。**

Rafi & Greenland BMC Med Res Methodol 2020より



訳註: 上の図はこの翻訳の原スライド (https://biostatistics.ucdavis.edu/sites/g/files/dgvnsk4966/files/inline-files/Greenland.Advancing%20statistics%20reform%2C%20part%204.Slides%201-110%2C%2001%20June%202022.pdf)のp.88より。出版バージョン

- Zad Rafi and Sander Greenland, Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise, 2020, https://doi.org/10.1186/s12874-020-01105-9 (https://doi.org/10.1186/s12874-020-01105-9)

の図2が上の図に対応しているが、その内容は少し違っている。

P値をS値(意外度)に変換して、
検定統計量が提供する証拠の強さを測れ(gauge)

- フィッシャー流のP値の扱いにおける中心的な要素は、異なる設定や検定においても、仮説やモデルに**反対する**証拠の共通尺度を提供することである。
- この尺度を日常的な感覚で表すために、任意のP値は独立で「公平な(fair)」(表が出る確率=1/2の)コイントスをs回行ったときに全て表が出る確率$1/2^s$と比較できる。
- P値 p が与えられたとき、そのpの値を与える表が連続する回数sを求める…

訳注: “S-value”は「S値」と訳し、“surprisal”は「意外度」と訳した。

- s回のトスで全て表が出れば $p = 1/2^s$
- sについて解くと $\mathbf{s = \log_2(1/p) = -\log_2(p)}$ となるので、
- $p = 1/2^4 = 0.0625$ は s = 4回のトスで4回表
- $p = 1/2^5 = 0.0313$ は s = 5回のトスで5回表
- $p = 0.04 = 1/2^{4.6}$ は $s = -\log_2(.04) = 4.6$ となる。

このようにP値に関する $p = 0.04 = 1/2^{4.6}$ は、約4～5回連続で表が出たことが「トス達は独立で表の出る確率は1/2以下であるという仮説の否定に与えるのと同じ程度の証拠を、P値 p を計算するときに使ったモデルの否定に与えることになる。

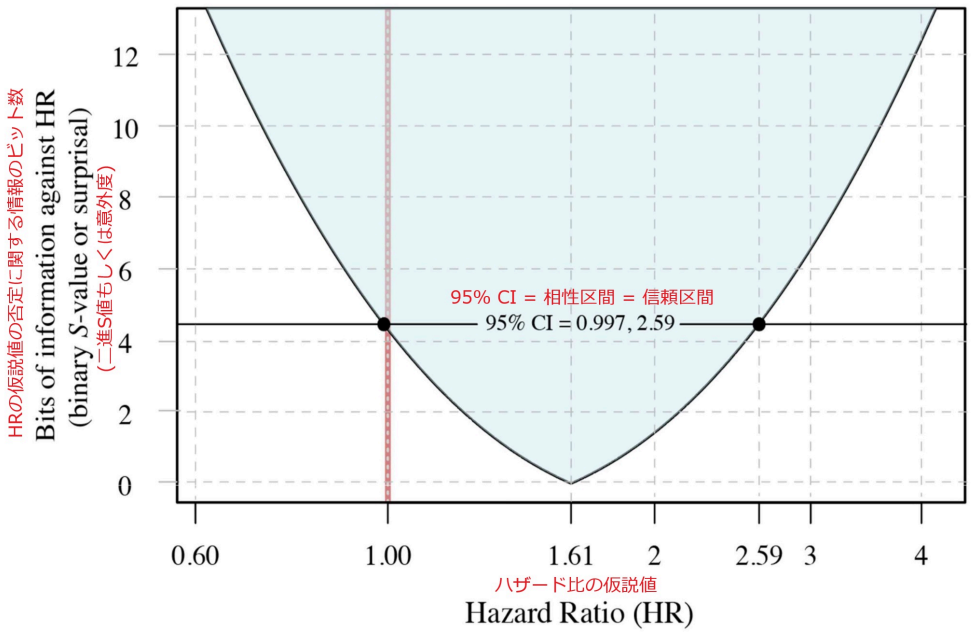
- **s = –log₂(0.05) = 4.3 ≈ 4回のトスで4回表**
- **s = –log₂(0.005) = 7.6 ≈ 7回のトスで7回表から8回のトスで8回表**

訳注: P値 p に対応するS値 s = −log₂(p) の差が1未満であることは差が非常に地位愛ことを意味する(1ビット未満の差)。だから、S値については四捨五入で整数に丸めても大した違いがないと考えられる。この立場はP値が5%未満であるか否かは本質的な大きな違いだと勘違いしている人達には受け入れ難いかもしれない。しかし、P値が4.9%と5.1%の場合で大きな違いがあると考えることが馬鹿げているだけではなく、P値が3%と6%の場合(それぞれのS値は約4と約5)を非常に大きな違いだと考えることも合理的ではない(1ビットの違いしかない)。

- **二進S値 (意外度、対数価値) s は、P値 p が p を計算するために使用されたモデルの否定について供給する情報を測る。**
- **s の単位は**シャノンまたは**ビットと呼ばれる。**
- **P値のスケールは高度に非線形である:** モデルの否定に関する情報の観点から、**0.001と0.05の差は大きいが、0.95と0.999の差は**、同じ距離だけ離れているにもかかわらず、**些細(trivial)である。**
- **S値はそれらの情報の差を示す:** −log₂(**0.001**) = 10, −log₂(**0.05**) = 4.3, Δ = **5.7 ビット**; −log₂(**0.95**) = 0.07, −log₂(**0.999**) = 0.01, Δ = **0.06** ビット

訳注: surprisal, logworthをそれぞれ「意外度」「対数価値」と訳した。

Rafi & Greenland http://arxiv.org/abs/1909.08579 (http://arxiv.org/abs/1909.08579) より



訳註: 上の図はこの翻訳の原スライド (https://biostatistics.ucdavis.edu/sites/g/files/dgvnsk4966/files/inline-files/Greenland.Advancing%20statistics%20reform%2C%20part%204.Slides%201-110%2C%2001%20June%202022.pdf)のp.92より。出版バージョン

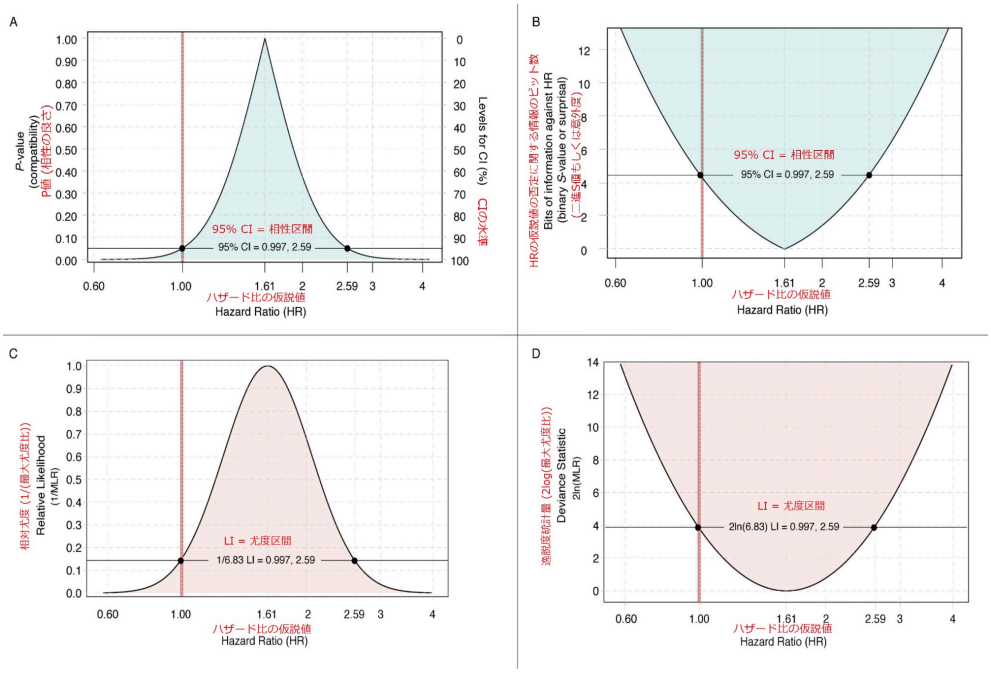
- Zad Rafi and Sander Greenland, Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise, 2020, https://doi.org/10.1186/s12874-020-01105-9 (https://doi.org/10.1186/s12874-020-01105-9)

の図3が上の図に対応しているが、その内容は少し違っている。

- S値は、理論家がP値に含まれる証拠や情報を評価し、対立仮説の下での検定の挙動を調べる必要があった1950年代以降、繰り返し再浮上している。

- **S値は1をはるかに超える範囲に及ぶため、ベイズ的な確率と混同しにくい。**
- **S値は事前分布を必要としない。しかし、次の方法でS値に事前分布を組み込むことができる:** 事前分布をパラメータサンプリング分布 (「ランダム効果」モデル) として扱う複合サンプリングモデルの適合度検定からpを計算する。尤度比との関係…

Rafi & Greenland BMC Med Res Methodol 2020より



訳註: 上の図はこの翻訳の原スライド (https://biostatistics.ucdavis.edu/sites/g/files/dgvnsk4966/files/inline-files/Greenland.Advancing%20statistics%20reform%2C%20part%204.Slides%201-110%2C%2001%20June%202022.pdf)のp.94より。出版バージョン

- Zad Rafi and Sander Greenland, Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise, 2020, https://doi.org/10.1186/s12874-020-01105-9 (https://doi.org/10.1186/s12874-020-01105-9)

の図1,2,S1,S2が上の図に対応しているが、その内容は少し違っている。図S1,S2は上の論文のAdditional File 3, 4にあるが、 https://arxiv.org/abs/1909.08579 (https://arxiv.org/abs/1909.08579) の第10節にもある。

図S1,S2には以下のようなキャプションが付けられている。

> **図S1：ハザード比の範囲に対する相対尤度**　図2(P値関数)に対応する相対尤度関数を示す。1/6.83 の尤度区間(LI)も描かれており、これは95%相性区間に対応する。Brown et al. [34] の結果から算出。**MLR** = Maximum-Likelihood Ratio (最大尤度比)。**HR** = 1 は関連なしを表す。

> **図S2：ハザード比の範囲に対する逸脱統計量**　図3(S値関数)に対応する逸脱度関数を示す。また、尤度区間(LI)も描かれており、これは95%相性区間に対応する。Brown et al. [34] の結果から算出。**MLR** = Maximum-Likelihood Ratio (最大尤度比)。**HR** = 1 は関連なしを表す。

私の見解に関する背景とさらなる参考文献
(以下のリンク先で公開アクセス可能なはずです)

- Greenland S. For and against methodology: Some perspectives on recent causal and statistical inference debates. Eur J Epidemiol, 2017;32:3-20.

https://link.springer.com/article/10.1007%2Fs10654-017-0230-6 (https://link.springer.com/article/10.1007%2Fs10654-017-0230-6)
- Greenland S. The need for cognitive science in methodology. Am J Epidemiol 2017;186:639-645. https://academic.oup.com/aje/article/186/6/639/3886035 (https://academic.oup.com/aje/article/186/6/639/3886035)
- Greenland S. The causal foundations of applied probability and statistics. In Dechter R, Halpern J, Geffner H, eds. Probabilistic and Causal Inference: The Works of Judea Pearl. ACM Books 2022; 36: 605-624, https://arxiv.org/abs/2011.02677 (https://arxiv.org/abs/2011.02677) (version with corrections)
- Greenland S. Analysis goals, error-cost sensitivity, and analysis hacking: essential considerations in hypothesis testing and multiple comparisons. Ped Perinatal Epidemiol 2021;35:8-23. https://doi.org/10.1111/ppe.12711 (https://doi.org/10.1111/ppe.12711)
- Greenland S. Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. Am Stat 2019; 73: 106-114. http://www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1529625 (http://www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1529625)

学生、研究者、編集者、教員に向けた学習のための文献

- Greenland S, Senn SJ, Rothman KJ, Carlin JC, Poole C, Goodman SN, Altman DG. Statistical tests, confidence intervals, and power: A guide to misinterpretations. The American Statistician 2016;70 suppl. 1, https://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_fil (https://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_fi
- Amrhein V, Greenland S, McShane B. Retire statistical significance. Nature 2019;567:305-307. https://www.nature.com/articles/d41586-019-00857-9 (https://www.nature.com/articles/d41586-019-00857-9)
- Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics. The American Statistician 2019;73 suppl 1:262-270. www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1543137 (http://www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1543137)
- Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. BMC Medical Research Methodology 2020;20:244 https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01105-9 (https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01105-9)
- Greenland S, Rafi Z. To aid scientific inference, emphasize unconditional descriptions of statistics. 2021, http://arxiv.org/abs/1909.08583 (http://arxiv.org/abs/1909.08583)

- **「Pearlのテーゼ」: 1900年頃、科学と統計学は、因果性を理論から排除し、さらには禁止しようと試みるという深刻な過ちを犯した。**
- 動機: 物理法則における時間対称性。
- しかしそれは、熱力学において現れる非対称性や、情報伝達の因果構造 (エネルギーの流れ、通信、因果の最大速度としての c) を見落としていた。
- 因果(パス)図式と潜在的結果モデルは1920年頃から存在するが、1990年頃まで完全に発展し、広く普及し始めることはなかった。
- **それらは基礎統計学に統合されるべきである！**

認知的盲目の図式の例(graphical example):
観察研究における因果的ゼロ仮説を擁護するための
節約性の誤謬(parsimony fallacy)

- 図式は、質的な性格をもつため、バイアスと分散のトレードオフについては何も語らない。そのため、純粋な予測モデルや潜在結果モデルに限定されている人達からはしばしば退けられる。
- しかし図式は、因果効果を評価するうえで統計的規準が不十分であることを示している。なぜなら**因果は確率だけでは捉えることのできない制約を含んでいるか**

らである。

訳註:

- 関数のグラフの意味での“graph”ではなく、変数記号のあいだに矢線を描いたものという意味での“graph”を「図式」と訳した。数学におけるグラフ理論の意味でのグラフは後者の意味である。
- それに伴い、“graphical example”を「図式の例」と訳した。
- 因果的ゼロ仮説(causal null)は「特定の矢線が表す因果効果はゼロである」の型の仮説のことである。効果ゼロを意味する矢線は因果図式中で省略される。
- “potential outcome”は「潜在結果」と訳した。

図式における矢線の欠如は、
因果的ゼロ仮説が仮定されていることを意味する

「ソフト」サイエンスでは、「効果なし」仮説と、その周囲の重要な効果を含む区間に含まれる対立仮説を区別することはほとんどできない。

- 技術的には、不連続分布(質量の集中をもつ分布)とそれを近似する連続分布の効果的な区別を実証的に (empirically)遂行することは不可能である。さらに…
- **連続性を点質量で置き換える近似誤差は、因果ネットワークを通じて増幅され、巨大な誤差となり得る。**

訳注: 確率論で「質量」は「一点に集中している確率」を意味する場合がある。

皮肉なことに、節約性(parsimony)を理由に
特定の効果の存在を否定する人々にとって、
ゼロ仮説が非実験的な観察に対する最も節約的な
因果的説明であることはほとんどない。実際、

- 何らかの関連が存在するとき、**「効果なし」というゼロ仮説は節約的ではない。**なぜなら、ゼロ仮説の下では、その関連を説明するために間接的な説明が必要となり、間接的な説明は直接的な因果関係よりも因果的により複雑になるからである。

訳注: 関連(association)と因果(causation)は異なる概念であることに注意せよ。しかし、関連が因果効果ではないと主張するためには、交絡の存在などのより複雑な前提が必要になる。

因果的節約性(causal parsimony)を観察された(ノンパラメトリックな)データ分布と相性が良い最も単純な因果図式を好むことだと定義したとする。このとき、

- 存在する関連を生み出すより複雑なメカニズムの体系に頼らない限り、報告された効果を却下する根拠は存在しない: そのために必要な因果図式はより多くの矢線とより大きな効果を要求する。

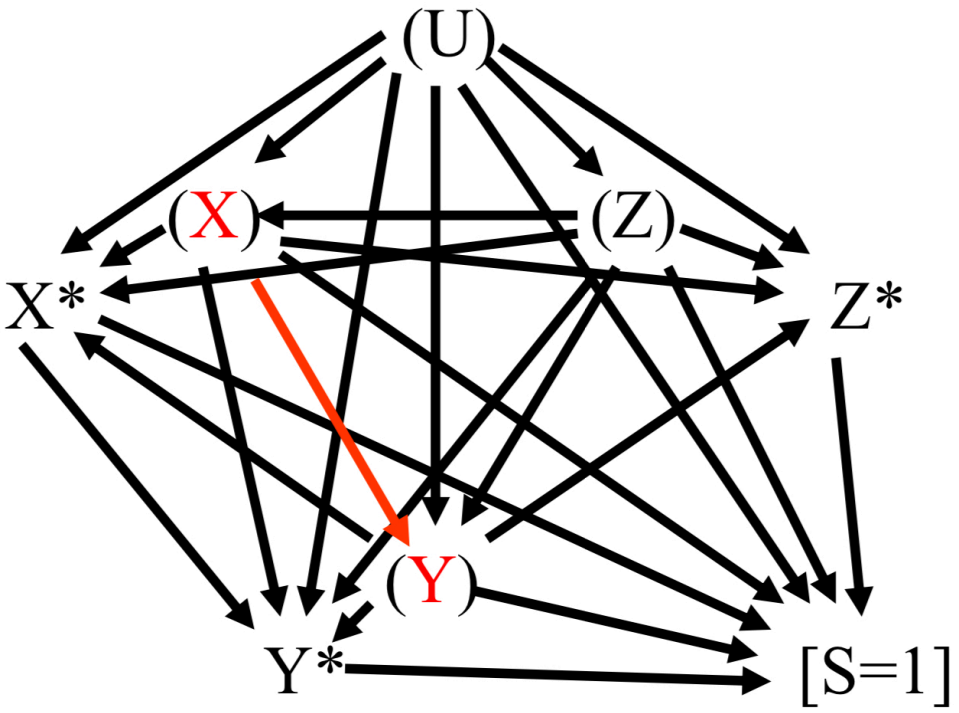
考察: X-Yの関連が観察された場合の最も単純な単一の説明は何か？:

- (a) 単純な交絡: X←C→Y
- (b) 単純な選択バイアス: X→[X]←Y
- (c) 差異的誤差: X→X*←Y もしくは X→Y←*Y*
  (*X*もしくは*Y*が誤差を伴う*X*\もしくはY*として観察される)
- (d) 単純なランダム誤差: X　Y←ε
- **(e) 単純な因果効果: X→Y**

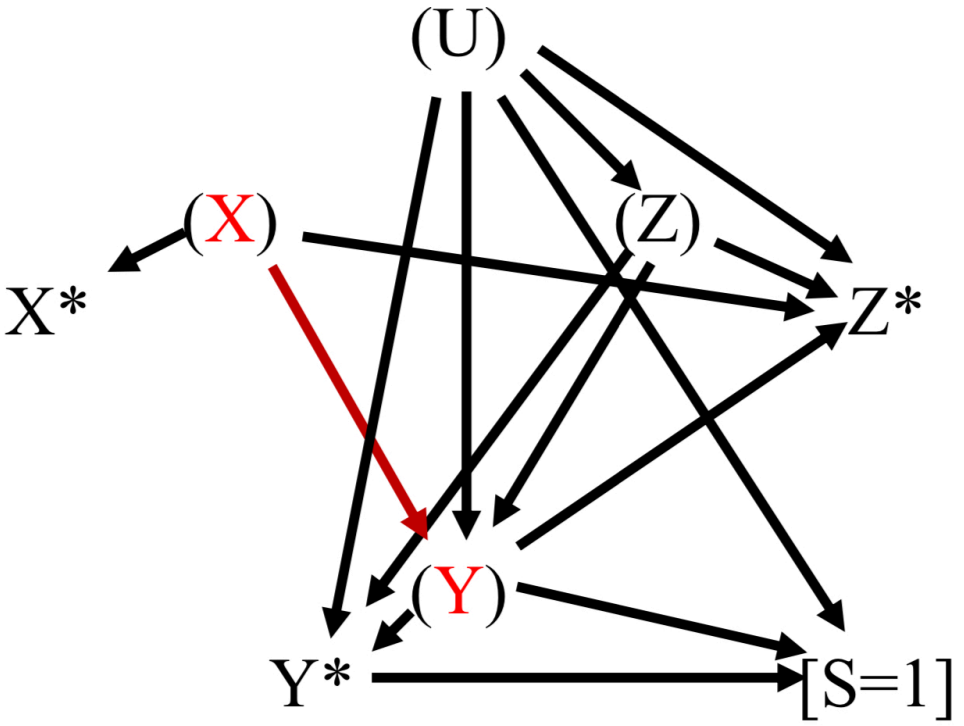
ゼロ仮説達(a),(b),(c),(d)は、単純な因果効果の(e)と比較して、余分な変数(ノード)や効果(矢線)を必要とする。

- 要するに、もしも何らかの関連が観察されたならば(その関連が「非有意」と宣言される範囲内にあるかどうかにかかわらず)、**ゼロ仮説を維持するためにはその関連の代替的な説明が必要となる。**
- その代替的な説明は、複雑性を説明の因果図式(cDAG)に必要な最小の変数と矢線の数として定義するならば、直接的な因果的説明(因果的ゼロ仮説の棄却)よりも常に複雑になる。

複雑な観察データの現実: XがYに及ぼすいかなる影響も、交絡、選択バイアス、測定誤差などのバイアス源の網の中に埋もれている:



X*-Y*の関連と隠れた変数達を含む最も単純で現実的なDAG達の中には、XとYがX→Yを通してのみd接続されているものが含まれる。例えば次のような場合である。



- 制御されていないバイアスが存在しない(XからYへの非因果的な開いたパスが存在しない)という仮説は、XからYへの観察された関係について提供できる最も節約的な説明である。
- しかし、ゼロ仮説を擁護するために節約性を持ち出す人達は、バイアスを生じさせる(開いた非有向)パスに節約性を適用せず、より複雑な代替案に絶対の自信を持っている。
- この行動は(特定のもしくは一般的な)ゼロ仮説への隠れた価値バイアスに影響された認知的な錯覚を露呈させている。

妥当な反対意見:
本質的な複雑さを反映できない場合には
節約性は誤解を招く

ニール・ドグラース・タイソンの言葉を言い換え: 「自然はあなたのために単純である義務を負っていない」

トウェイン曰く: 「あなたを困らせるのは、あなたがモデル化したものではなく、**あなたがモデル化しなかったけど、そこにあるものだ**。」

- 図式中の2つの変数間に矢線がないことはゼロ仮説を意味する。

矢線やノードの削除を**因果的に**保証するものは何か？

答え: 因果デザインによる削除の強制
— 例: コホートマッチング(ブロッキング)、ランダム化。

- Xがランダム化されていれば、Xへの矢線を削除できる。
- 無作為抽出が行われていれば、Sへの矢線を削除できる。
  しかし、観察研究の定義より、
- 研究処置 X はランダム化されていない。
  さらに、健康科学の現実においては、
- 選択・参加 S は無作為ではない。

ランダム化されていないこと = 「客観的」統計が存在しないことである。あるのは「このモデルの下では…」という条件付きの言明のみ。

- 「ソフトサイエンス」では、ゼロ仮説の近くに固く集中した事前分布が実際の証拠に基づくことはほとんどない。ある特定の設定では多少の支持を得るかもしれない(例: ゲノミクス)。
- もしもすべての因果パスがランダムウォークであれば、実際の効果はゼロの近くに集まるかもしれず、ほとんどの効果サイズを「重要でない」ものにするだろう… しかし、
- 「重要性」の判定は価値判断を伴う(value laden)。効果が正確にゼロだと宣言してしまうことはこの問題を覆い隠す。

- 連続性がある場合には「偽陽性 (false positive)」はほぼ存在しないとみなされる。なぜなら、ほとんどすべての**関連(associations)**は非ゼロ(「真陽性 (true positive)」)だからである。
- 「偽陽性問題」は、効果をいつ枝刈り(prune)または無視するかという問題の歪んだ過剰な単純化であり、その意思決定には損失(ペナルティ)関数が必要である。
- 効果的な枝刈り(pruning)アルゴリズムは事前分布の連続性を保持できる。例: 絶対距離(LASSO、ラプラス型)を二乗距離(ガウス型)ペナルティの代わりに使用すること。
