若手のための Multivariate 研究会(M 研)　07/08/04

Kline, R. B.(2004)　Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research　APA

Chapter1：Changing Times

Rep.脇本竜太郎[1]

＊本書のウェブサイト(http://www.apa.org/books/resources/kline)で，発展的事項についての補足，解答付の練習問題，講師・学生のための資料，関連ウェブサイトへのリンクが閲覧できる。

本書の歴史的背景

■1996 年，APA の科学問題委員会(Board for scientific affairs)が TFSI(Task Force on Statistical Inference；統計的推論に関する特務委員会)を召集，長い間続いている統計的有意性検定に関する議論に答え，それに代わる手法を明確にすることを求める。

←一部の人は TFSI が心理学論文誌での統計的検定の禁止を勧告することを期待

…このような禁止論については Harris(1997a；Psychological Science)，McLean & Kaufman(1998；Research in the Schools)，Harlow et al.(1997)編の”What if there were no significance tests?”等でも議論されている。

…また，統計的検定に代わる方法の問題点については， Journal of Experimental Education の特集号(B.Thompson,1993a)で議論されている。

■このような禁止論の背景には，有意性検定への幻滅が来るところまで来たことを反映。

・実際，このような議論は年を追うごとに激しくなり，心理学から野生生物科学まで領域を超えて行われるようになっている。

・このような議論は，心理学や関連領域の論文誌で統計的検定の結果が報告されるのが普通である者にとっては，驚くべきもの。

・しかしながら，1990 年代後半までには行動科学の分野から統計的検定の限界を唱える声は無視できないほどに多くなり始めている。

→このことは，約 2 ダースにおよぶ論文誌(心理，教育，カウンセリング等)で，投稿原稿に効果量の記載を求めていることからも明らかである。

…統計的検定のみでは不十分である，という投稿者への強いメッセージ

・このうち 2 つは，会員数 50000 人を超える学会(American Counseling Association, Council for Exceptional Childern)の機関(&旗艦)紙

[1] 日本学術振興会・東京大学

・また,APAJournal のうち，このような要求を行っている最新のものは Journal of Educational Psychology である。

・このような著名な論文誌の編集方針は変化を先導するものとなりうる。

Kaufman(1998)：統計的検定についての議論は我々の世代の主な方法論的問題と指摘

→変化の時がやってきている。

Goals and Plan of the Book

■本書の目的：読者が行動科学領域でのデータ分析・知見の報告の仕方についての今まさに生じている変化を追うための一助となること。

想定される読者：教育者，応用研究者，論文原稿査読者，心理学の大学院生・学部生

・多くの者が，伝統的なデータ分析の手法の訓練を受けている者…つまり，統計的検定が主要な分析方法だと思っていることを想定している。

・学生読者は，視野や技能が狭くなっていないという意味でいくらか有利であろう。また，既に統計的検定の限界を確信している者は，自己の視点を強化する有用な議論を本書に見出すだろう。また，そのような視点を持っていない者にとっても，熟慮すべき興味深い点を本書に見出すことが期待される。

■本書では，心理学領域の者である我々が，現在の統計的検定の使い方を改めるべきか否かという点については議論しない。

・Tyron(2001)50 年以上にわたる統計的検定を正しく用いることの利益を協調して誤用を正そうとした試みが成功していないことを指摘

⇔本書では，すでに当該領域での発展が検定の役割を縮小させる方向を指し示していると想定。これに基づいて，本書の目的は以下の点について読者の理解を支援すること

(a)統計的検定に関する論争と限界

(b)複数の統計的検定に対する代替案の強みと弱み

(c)統計的検定の役割縮小に関連した方法(メタ分析等)

＊(b)と(c)については，効果量の推定が最も重要

＊他に，本書では区間推定，特に観測効果量に基づいた信頼区間の算出にも焦点を当てる。

■本書の構成

Part Ⅰ：基本的な概念と統計的検定に関する議論のまとめ

Chap.2：信頼区間と有意性検定の根底にあるサンプリングと推測の原理

Chap.3：仮説評価の主要な手段として統計的検定を使い続けることに対する議論

*この議論では統計的検定そのものに問題があるのではなく，それが実際に行っていることがおそらくほとんどの行動科学研究について不適であると想定

*また，統計的検定への捉われが心理学の発展を阻害していることも議論される

PartⅡ：2 群以上を比較する研究における効果量推定

Chap.4：効果量推定の原理的説明，連続的結果変数に対する基本的なパラメトリックな効果量指標の紹介。実際的有意性(substantive significance；理論的，臨床的，実際的)と統計的有意性に関する重要な問題についての議論

Chap.5：カテゴリカルな結果変数について 2 群を比較する際のノンパラメトリックな効果量指標の紹介

Chap.6・7：3 水準以上の一要因デザイン，2 つ以上の独立変数を持つ要因デザイン

＊Chap4~7 では，特に必要不可欠な方程式のみを呈示。これらのうちには元データに適用されるものだけでなく，2 次分析に手ごろなものも含まれる。
その他の効果量指標算出方法については，Cooper & Hedges(1994)などメタ分析に関するテクニカル本で参照可能。

PartⅢ：社会科学領域のデータ分析手法の再構成に関する議論

Chap.8：再現性(replication)とメタ分析

Chap.9：統計的リサンプリング(ブートストラップ法など)とベイズ推測

Retrospective

Hybrif Logic of Statistical tests(1920-1960)

■帰無仮説検定(Null Hypothesis Significance Testing；NHST)の要素は 1700 年代に科学論文に登場…しかし，それら要素は系統的な方法としてまとめられていない

・現在の NHST の形は，R.Fischer(1925)のアプローチと Neyman –Pearson アプローチ(Neyman & E.S.Pearson,1933)の合成物

■Fischer のアプローチ

・帰無仮説のみを問題とし，その下での条件付確率を統計的検定によって求める

・検定によって算出された確率=p 値…”p-value approach”と呼ばれる

・5%水準，1%水準という伝統的基準は Fischer が定めたと一般的に言われるが，Fischer 自身はどの研究でもその基準を使え，などということは言っていないようである。

■Neyman-Pearson アプローチ

・Fischer のモデルに対立仮説を加えた

・片側，両側棄却域の特定

・全研究で統一した確率(限界水準，α)を用いる
…”fixed p approach”，現在 5%や 1%を頑なに用いる慣行の原因

・検定力，第 1 種・第 2 種の誤りといった概念的枠組みを与える

…検定力分析は Fischer のモデルではなく，Neyman-Pearson モデルに基づいている。

■2 つの立場は激しい論争を展開

・1930~50 年ころに，上記 3 者以外の統計家等により 2 つのモデルが統合され，現在の NHST が誕生

…hybrid logic of scientific inference(Gigerenzer,1993),Intro Stats(Dixson & O'Reilly,1999)

・現在の NHST について

(a)Fischer，Neyman，E.S.Pearson の何れにも受け入れられなかったと思われる

(b)合成物としての性質が，統計的検定の結果が意味するところについて混乱が存在する原因であることが指摘されている

Institutionalization of the "Intro Stats" Method(1940-1960)

■1940 年代以前：検定は公刊論文ではほとんど使われず，標準外の方法で様々な記述統計や初歩的なテスト統計を使用

■1940～1960："inference revolution"(Gigerenzer & Murray,1987)

・Intro Stats が"the"仮説検定の方法として，教科書，大学のカリキュラム，論文編集・査読活動に適用されるようになる

←変化の 2 つの要因

・シングルケース研究からの脱却

・科学における"probabilistic revolution"：主観的事象をよりよく理解するために，量子論や遺伝学領域での非決定論を紹介

→しかし，心理学領域において，これは NHST を通して推測のプロセスを機械化させるために使用された

■Intro Stats の導入後，心理学および関連領域において，統計的検定を報告する論文の割合が急増(Figure1.1 参照)

*ランダムに選択された 12 の APA 誌の，1911-1998 間の約 8000 本の論文が対象。

■Gigerenzer(1993)が指摘する，NHST 制度化の利点

・論文編集者の審査の簡易化

・NHST は機械的に適用されるため，推論から主観的判断を除いているように見えた

(この客観性が実際よりもそう見えているだけだ，という点については後に触れる)

・NHST が行動科学者の共通言語，あるいは大きな研究事業の同じ構成員としてのアイデンティティの礎となった

いかなる方法でも，それが教義にまで高められてしまうことにはコストが存在する。次節ではそのいくつかについて触れる

Increasing Criticism of Statistical Tests(1940-present)

■NHST に批判的な著作の数は，1940 年代以来急激に増えている

・Figure1.2：Andersen et al.(2000)が，生態学，医学，ビジネス・経済，統計学，社会化科学分野から統計的検定に批判的な論文の本数を年代別にまとめたもの。

■以下 6 点は，行動科学領域での統計的検定の継続的使用に対する批判的意見をまとめたもの。詳しくは後述する。

1.p 値に対する誤解(サンプリングエラー，再現性，いずれかの仮説が正しい確率等)は，ユーザー側のみの問題でなく，近代の NHST の論理的裏づけが一貫していないため生じている。

2.統計的検定が意味するところに対する誤信念が，行動科学領域の研究の発展を阻害している。これは自然科学比べてより強い再現性の伝統が築かれていないこと，研究の関連性の欠如，研究労力と資源を無駄にしていることからも明らか。

…このような問題は，統計的検定そのものよりも過剰な使用によって生じると考えられている。

3.多くの研究で報告される p 値は，特に怪しい帰無仮説を検定していたり，分布の仮定が満たされていない場合，信用できない。信用できない p 値に基づく議論は，やはり信用できない。

4.統計的検定がもたらす情報は極めて特定的で，そのため概して研究者が本当に知りたいことを伝えない。

5.統計的検定は効果の大きさや，それが理論的，実際的，臨床的に意味があるかどうかを直接には示さない。効果量の大きさ，実質的有意性(substantive significance)，再現可能性こそが我々の知りたいことである。

6.統計的検定に様々な問題があれども，魔法のような代替策はない。それゆえ，代替策自体も無批判に支持されるべきではない。

…個別の研究での効果量の報告や信頼区間の算出，またメタ分析によるそれらの統合にも，それぞれの問題がある。

The Failure of Early “Suggestions” to Report Effect Sizes

■統計的検定の限界の埋め合わせをする 1 つの方法は，効果量指標等補足的情報を報告すること

・効果量指標自体は新しいものではない;K,Pearson の 1900 年代初期の著作に既に登場，また $\hat{\eta}^2$ (あるいは相関比)は Fischer によるものとされている。

■APA の投稿マニュアル第 4 版では，著者に統計的検定の結果と共に，効果量を報告することを奨励(要求はしない)→しかしうまくいかず

・Kirk(1996)：4 つの APA 論文誌の 1995 巻の論文を検証

→効果量を報告した実証論文の割合は，4 誌で 12～77%。しかし，77%の雑誌はもともと回帰分析が主に用いられており，自動的に相関効果量(R^2)が算出されている。これを除いた 3 誌の平均は 25%。

→また，著者は必ずしも効果量の解釈を行っていない。

・Finch et al.(2001)：Journal of Applied Psychology 誌において，1940 年から 1999 年まで，統計的検定の結果の報告のあり方が変わっていないことを報告。

＊高価ではないパーソナルコンピューターが多くの洗練された統計的手法の使用を可能にしたというのに。

The Rise of Meta-Analysis and Meta-Analytic Thinking

■1970 年代の登場以来，メタ分析は研究を統合する重要なツールとなっている。

→行動科学分野では，標準化効果量の中心傾向・変動を推定するために使用される

→統計的有意性でなく効果量への着目は，メタ分析論文の読者に統計的検定の限界を，既成概念に捉われず考えることを促す

→現在は，個々の研究での帰無仮説の棄却/保持に基づいた結論が，メタ分析により誤りだったことが明らかになった例が複数存在。

■メタ分析使用の増加は，メタ分析的思考を促進している。この動きはさらに支配的になりつつあるようである。

・メタ分析的思考の特徴(Cumming & Finch, 2001；B. Thompson, 2002b)

①先行研究の結果の正しい理解が必要不可欠であるとみなす

②研究者は，自分の研究が先行研究全体に控えめ(modest)な貢献をすると考えるべき

③研究者は，結果を将来のメタ分析に取り入れやすいように報告すべき

④新しい結果を，先行研究の効果量と直接比較して回顧的に解釈することが必要

・これらは統計的検定を第一の推論ツールとして使用することとは相容れない。

Report of the TFSI and the APA's Fifth Edition of the Publication Manual(1999-present)

■TFSI は当時発行予定だった APA マニュアル第 5 版に複数の勧告を行った。

・主要なものの一部

①最小限に十分な分析を行う

②意味を理解しないまま，コンピュータ出力から結果を報告しない

③母集団効果量に対する想定，サンプルサイズ，当該研究の検定力の演繹的な推定量の

背後にある測定法 [2]を報告する。観測検定力を報告する代わりに，観測された結果の信頼区間を使用する

④主要な結果，あるいは p 値が登場する場合にはいつでも効果量を報告する

⑤観測効果量の信頼区間を報告する

⑥データが統計的想定に適合していることを，合理的範囲で保証する

・しかし，TFSI は統計的検定の禁止は勧告しなかった。

…濫用を抑制する方法としては，禁止は極端すぎるであろうと看做されたため

■APA マニュアルの第 5 版も同様の立場に立っている。

・統計的検定に関する議論の存在を認めつつ，その議論の解決はマニュアルの役割ではない(p21-22)との見解。

・統計的検定結果の完全な報告を奨励：テスト統計量，自由度，限界水準もしくは有意水準。

・その他の統計的検定に関する勧告は以下(p21-26)

①十分な記述統計を報告：平均，分散，群の人数，要因研究でのプールされた級内分散 - 共分散行列，回帰分析の相関行列→後のメタ分析，2 次分析に必要

②効果量はほとんど全ての場合に報告されるべき(p25)とし,効果量指標の例が列挙されている。また，効果量を報告しないことが研究の欠陥の例として挙げられる。しかし，報告は要求されていない。

③信頼区間の使用を強く推奨。しかし，やはり要求されていない。

■予測できたことではあるが，皆が TFSI の報告や APA マニュアル第 5 版を好ましく思っているわけではない。

・Sohn(2000)：心理学研究間の連関性を改善するような，統計実践の変化のための明確なガイドラインがないことを嘆く

・Finch et al.(2001)：TFSI の報告を歓迎しつつ，現行の APA マニュアルの曖昧な勧告を International Committee of Medical Journal Editors(1997)による生化学領域の論文ガイドラインと対比

・Kirk(2001)：TFSI の報告は歓迎。一方で APA マニュアル第 6 版では TFSI の勧告についてより詳しいセクションを設けることを示唆

■なぜ第 5 版では効果量の報告を要求しなかったのか？

・Fidler(2002)の当事者(principals)へのインタビュー

：複雑な反復測定デザイン，多変量デザイン等，効果量の算出が困難あるいは不可能な

[2] Measurement behind a priori estimates of the statistical power of the study。想定した検定力がどうやって導かれたかということを説明しろということ？

場合が存在するため

→しかし，(それでも)ほとんどの行動科学の研究では効果量が算出できる。

→現状効果量推定ができないようなデザインに関しても，研究が進んでいる。

PROSPECTIVE

■ここまで述べた出来事は，統計的検定の役割がどんどん小さくなる未来を予測

…しかし，それは一朝一夕には起こらないだろう。検定結果を報告しなければ論文がリジェクトされてしまうだろうから。

■それでも，解釈の際の検定の重み付けは減少させて然るべき。具体的勧めとしては

①統計的に有意な結果を特に有益(注目すべきだとか再現されるとか)だと考えるべきでない。

②非有意な結果を割り引いて考えるべきでない。

例えば，帰無仮説が棄却できないからといって母集団における効果がゼロだと結論すべきでない。

③効果量は常に報告されるべきであるし，信頼区間も可能な場合はいつでも算出されるべき。但し，有意な結果について効果量を報告するというのではなく（これでは結局元の木阿弥で検定中心になってしまう），効果量の実質的有意性を解釈・評価する。

■その他の勧めは後の章で触れるが，その中には全く統計的検定の使用を含まないものもある。

→著者者とその仲間が提唱する行動科学の将来の展望と整合

・検定を判断の主たる基準と捉えない

・等質性検定や推測的信頼区間の算出等，特定の場合時にだけ NHST を使用

・社会科学が自然科学に近づく

…効果の方向と大きさを報告し，それが再現されるか判断し，さらに理論的，臨床的，実際的有意性から判断する(ただ単に統計的有意性からではなく)

VOICES FROM THE FUTURE

■前述の通り，若い研究者は方法論が凝り固まっていないので，データ分析・推測手法の変革の声により容易に答えることができる。しかしながら,変革の手綱はいくつかの雑誌の編集者(典型的に業績も経験も豊富)および当書に引用されている多くの著作の著者らの手に握られている。

■著者らの経験から，学生は変革の有望な担い手

・若い学生は統計的検定の限界について学ぶことに熱心

・適切な講義の下で，初級の統計の講義でも効果量や信頼区間等の発想について理解することができる。
・(著者の経験上)学部生に上記の概念を教えるほうが，入り組んだ統計的検定の話を教えるのよりも容易いし，再現性の概念などはもっと教えやすい。

■最近の講義の最終試験時に行った”What is the most important thing you learned in this class?”という問いへの学生の回答
・NHST が仮説検証の唯一の方法でもなければ，必ずしも最善な方法ではないということ
・ある発見が統計的に有意だということが，それが重要だとか信頼できるだとかいうことを意味しないということ
・サンプルサイズを十分に大きくすれば，全ての結果が統計的に有意になること。これは恐ろしい。
・統計的検定を過剰に強調している論文には懐疑的になること。
・統計的有意性は実用的有意性を意味しない。効果量，検定力，平均，標準偏差は研究報告に含まれるべきである。社会科学領域には統計的検定のよりよい理解が必要で，それが達成されれば我々はよりよい，より多くの情報に基づいた選択を行うことができる。

■今こそ統計的検定を乗り越え，他の古い慣行を脱して，我々の将来を築くべき時。我々は仮説検証の他の可能性―心理学および関連領域の研究により生産的な未来をもたらすものーを模索する必要がある。本書が読者をこの方向に喚起し，向かわせ，また支えるものとなることを期待する。

CONCLUSION

■第 1 章では，心理学における統計的検定にまつわる議論，およびその議論につながる出来事について概観した。
・これら歴史は，1999 年の TFSI 報告および 2001 年の APA マニュアル第 5 版の文脈を示すもの。
・また，統計的検定を使い続けることが行動科学において仮説検定，推測的判断を行う唯一の方法でないことを示唆

■本章で取り上げた点は，次章で扱う基本的な統計概念の概観の舞台を整えるものであり，それら概念は統計的検定の限界及び提案される代替案(区間推定や効果量推定)の特性を理解するのに重要なものである。