Kline R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research* APA.

CHAPTER3
WHAT'S WRONG WITH STATISTICAL TESTS___ AND WHERE WE GO FROM HERE

Rep.脇本 竜太郎[1]

■この章では，帰無仮説検定(null hypothesis significance testing;NHST)の問題について考察する。NHST を巡る議論を概観した上で，それらが NHST 利用の縮小あるいは放棄を支持することについて議論する。さらに，NHST の置き換えあるいは補完，また学術雑誌の編集方針や教育カリキュラムについての示唆を述べる。

NHST OUTCCOMES ARE INTERPRETED AS SOMETHING THEY ARE NOT

■よく発達したパターン認識能力ゆえ，人は時に，本来無作為なものにまで意味を見出してしまう。

・生まれ持っての性質と訓練ゆえに科学者は極めてパターン認識が得意。それゆえ無意味に意味を見出す過ちを犯しやすい。

・NHST の場合も同様で，それに関するよくある誤解は，それが推論できる以上のことを誇張するというタイプのものを含んでいる。そのような誤った推論は，行動研究の進歩を妨げるような迷走[2]の原因の１つ。

Misinterpretation of p values

■p 値とは，帰無仮説 H_0 が正しい場合の統計量の確率。p<.05 の正しい解釈は以下のようなもの。他の正しい解釈は表現は違っても，これらのバリエーションである。

1. H_0 が真である場合に，無作為な標本から観測された以上の極端な結果を得るオッズは 1 対 19 よりも小さい。
2. 観測された結果よりも，H_0 の下での標本の分布の平均値から遠いような検定統計量は 5%より少ない。
3. H_0 が真であり，研究が無限に繰り返されると仮定すると，5%以下の結果が，観測された結果よりも H_0 との不整合の程度が大きい。

→p<.05 から言えることはこれだけ。いかなる正しい p 値の定義も，$p(D|H_0)$と表記することにする。

■以下では p 値そのものに関する誤解を列挙。それらのいくつかは，条件付き確率であることを忘れている，または p 値，D，H_0 が示す事象を逆転してしまっていることから生じている。

【誤解①：p 値は，結果が標本誤差により生じている確率である】(odds-against-chance fantasy)

・p 値は，標本誤差が帰無仮説からの標本統計量の乖離を生ぜしめるという前提の下で算出される。つま

[1] 安田女子大学文学部心理学科

[2] cognitive misdirection の意訳。

り，検定を行う時点で標本誤差の生じる確率は 1．ゆえに，上記の考え方は誤り。

・この誤解は後述のものと合わせて，統計的検定が結果を偶然(帰無仮説保持)と本当の効果(帰無仮説の棄却)に二分するものだという誤解を説明するかもしれない。第Ⅰ種及び第Ⅱ種の誤りの影響で, NHST の結果に基づくいかなる判断も，誤りであるかもしれないのだ。

【誤解②：p 値は，結果が与えられた時の，帰無仮説が真である確率である】(inverse probability error)

・これは，p 値がデータの確率（$p(D|H_0)$）であることを忘れ，データが得られた場合の帰無仮説の事後確率（$p(H_0|D)$）だと勘違いしている。

・しかし，$p(H_0|D)$は研究者が本当に知りたいものかもしれない。ベイズの定理は，この事後確率が p 値と関連していることを示している。

$$p(H_0|D) = \frac{p(H_0)p(D|H_0)}{p(D)}$$

・不幸なことに，古典的な統計的検定を用いる者は事後確率について考えないし，強いられて推測してもそれを主観的だとみなすだろう。→事後確率も扱うベイジアン・アプローチについては 9 章で扱う。

【誤解③：帰無仮説が棄却された場合，p 値はその判断が誤りだったことの確率(つまりタイプⅠの誤りの確率)】

・タイプⅠの誤り　$\alpha = p(\text{reject } H_0|H_0)$。一方，上記誤解は $p(H_0|\text{ reject } H_0)$。$\alpha$ から後者を推定することは，一般的に可能でない。

・個々の研究での帰無仮説の棄却の判断は正誤のいずれかで，確率的な問題ではない。棄却の判断の正誤についての推測は十分な追試によってのみ可能となる。

【誤解④：1-p は，観測データが与えられた場合の対立仮説が真である確率】(validity fallacy)

・正しくは，1-p は，帰無仮説が真である場合に，観測データよりも極端でないデータが得られる確率。これは対立仮説の事後確率とは直接関係がない。

【誤解⑤：1-p は，条件が一定の場合に結果が再現される確率である。】（replication fallacy）

・もし将来の再現の確率を知ることができるのなら，とても便利だ。しかしながら，実際は p 値というのは，特定の仮説の下での特定のデータに関する確率である。

・Carver(1978)が指摘しているように，結果の再現は実験デザインと母集団で本当に効果が存在しているか否かの問題である。

・Greenwald et al.(1996)は, p 値が追試の統計的検定力と monotonically に関連することを示している。しかし，この関係は順序的かつ非線形で，p 値を直接それに変換することはできない。

Mistaken Conclusions After Making a Decision About theNull Hypothesis

■p 値に基づいて仮説保持・棄却の判断をした後にも，様々な誤った結論が見られる。その多くはなぜ間違っているか説明する必要すらないもの。

【誤解①：p 値は効果の大きさの数値的指標である】（magnitude fallacy）

・効果量推定のような他の分析を行わない限り，そんなことは言えない。なぜなら，p 値[3]はサンプルサイズと効果量双方を反映しているから。つまり，効果量が小さくてもサンプルサイズが大きければ統計的に有意になる。サンプルサイズが大きすぎる場合，統計的検定はサンプルサイズが大きいことを確証するというトートロジーに陥る。

【誤解②：帰無仮説の棄却は，対立仮説とその背後にある研究仮説の確証を意味する】（meaning fallacy）

・これは 2 つの概念的誤りを反映

-まず，単一の研究での H_0 の棄却は H_1 の確証を意味しない。

-次に，統計的な仮説 H_1 が正しいことは，背後にある実質的(substantive)仮説も正しいことを意味しない。

○Arbuthnot(1710)の事例：

-82 年分のロンドンの出生記録をしらべ，男児が女児よりも多いことを確認し，男女の出生数が等しいという帰無仮説($H_0: \pi = .50$)を棄却。

-しかし，彼の実質的仮説は「神の意志が，男性の死亡率の高さを補おうとして男児の出生数を多くしている」というもの。これは誤り。実際は，Y 染色体を持つ精子が X 染色体より泳ぐのが速く，卵に到達しやすいため。

・統計的仮説と実際的仮説の区別は極めて重要。それは抽象度だけでなく，H_0 棄却後に異なる示唆を与える。

- H_0 と H_1 が統計的仮説のみを反映しているのなら，H_0 棄却後には追試以外やることがない。

-しかし，H_1 が科学的仮説を表しているのであれば，まさに H_0 棄却後に作業がはじまる。

…H_1 と整合する他の実質的仮説との比較検討，等価なモデルの反証など

“strong inference”（Platt,1964）

【誤解③：帰無仮説を棄却できないということは，母集団効果量が 0 であることを意味する。】

・科学の基本的定義：証拠の不在は不在の証拠にはならない。

・また，帰無仮説を棄却できなかたのは，第Ⅱ種の誤りなのかもしれない。

-行動科学研究の検定力は相対的に低いので，そのような事態は少なくない。

【誤解④：$H_0: \mu_1 = \mu_0$ が棄却できないといことは，母集団が等価であることを意味する。】

・たとえば，効果のある確立された治療法と，より低コストな新しい治療法を比較するケースを考える。$H_0: \mu_1 = \mu_0$ が棄却されたからといって，この 2 つの治療法を等価と見なすことはできない。信頼性係数や集団の比率の場合に関しても同様。

・差異に関する証拠の不在は，等価性の証拠にならない。等価性検証のための正しい方法は後述する。

【誤解⑤：帰無仮説の棄却は，実験デザインの質を裏付ける】

・まずい実験デザインが人為的な効果によって H_0 棄却を導くことがある。また，よく統制された研究で

[3] より正確には検定統計量。

も，標本誤差によって第Ⅰ種の誤りを犯すことがある。

【誤解⑥：帰無仮説を棄却できなければ，研究は失敗である。】

・方法のまずさや検定力の低さが第Ⅱ種誤りを導くことがある。しかしながら同時に，H_0を棄却できないことはよい科学の産物であるかもしれない。

-追試で帰無仮説が繰り返し棄却できないことが，最初の研究の誤りを明らかにすることも。

【誤解⑦：H_0の棄却は，根底にある因果のメカニズムを特定したことを意味する。】

・単一の H_0 の棄却は統計的仮説 H_1 に代表される，想定された因果の効果を証明しない。

【誤解⑧：追試の失敗とは，H_0に関して研究間で同じ判断ができないことである】(reification fallacy)

・この考え方は，サンプルサイズや効果量，2つの研究の効果を無視している。

-1つ目の研究で有意になった平均値の差が，2つ目の研究ではサンプルサイズが小さいことによって有意にならない，という事態は生じる。

-研究間で H_0 に関する異なる判断が下される場合であっても，再現性について肯定的な証拠を得ることができる。

Widespread nature of Misinterpretations

■ここまでで述べた誤解は，研究者や教育者にも普通に見られる。

・Oakes(1986)：p<.01 の解釈の誤り(Table3.1)。正しいのは 7 のみ。

TABLE 3.1
Usually Adopted Interpretations of $p < .01$ by 70 Academic Psychologists

Statement	*f*	%
1. The null hypothesis is absolutely disproved.	1	1.4
2. The probability of the null hypothesis has been found.	32	45.7
3. The experimental hypothesis is absolutely proved.	2	2.9
4. The probability of the experimental hypothesis can be deduced.	30	42.9
5. The probability that the decision taken is wrong is known.	48	68.6
6. A replication has a .99 probability of being significant.	24	34.3
7. The probability of the data given the null hypothesis is known.	8	11.3

Note. From *Statistical Inference* (p. 81), by M. Oakes, 1986, New York: Wiley. Copyright 1986 by John Wiley and Sons. Reprinted with permission.

・Tversky & Kahneman (1971)：belief in the law of small numbers

-(a)サイズの小さいサンプルでも，母集団を代表している，(b)統計的に有意な結果は，サイズが半分の追試サンプルでも得られる，という信念

・Table3.1 に示されたような誤りは，出版された文献にも容易に見出すことができる

Cohen(1994)：(自分自身を含む)素晴らしい著者が犯した誤りを列挙

Dar et al (1994)：30 年間の期間で，最高峰のピア・レビュー誌に論文を発表した有名な精神療法の研

究者が犯したミスの指摘

・学部と大学院時代に統計の授業で NHST の勉強や教育に時間をかけているのにこのような事態が生じるのは，不可解。

-いくつかの点を指摘可能。まず，NHST は平明な推論システムではない(用語が特定的，フィッシャーとネイマン・ピアソンの理屈の組み合わせに内在的矛盾が存在)。また，人間は一般的に条件付き確率を伴う推論に弱い。

NHST DOES NOT TELL US WHAT WE REALLY WANT TO KNOW

■ここまで概観してきた誤解は，研究者が本当に知りたいことと関連している。例えばデータが与えられた時の

・H_0 と H_1 が正しい確率　$p(H_0 | D)$，$p(H_1 | D)$

・結果が再現される確率　$p(R | D)$

・H_0 の棄却が誤りである確率　$p(H_0 | Reject\ H_0)$

- これらの希望をすべて叶える方法はない(ベイジアンアプローチは別。これは 9 章で紹介)

- それでも，我々の知りたいことを教えうる方法がある。それは NHST ではなく，追試である。

NIL HYPOTHESIS ARE USUALLY FALSE

■社会科学で最もよく検定される仮説は，ゼロ仮説(nil hypothesis)。しかし，特に 0 が効果の完全な不在を意味する時, 母集団パラメタの値がちょうど 0 であるというのは極めてあり得なそうなこと。

・例えば何らかの変数同士の母集団相関 ρ には，ゼロ以外の値を仮定するのが現実的

…crud factor(Meehl, 1990)，ambient correlational noise (1968)

…また，同じ測定方法が用いられることで生じる common method variance により，相関は 0 から乖離してしまう(絶対値で.2~.3 程度)

■仮にゼロ仮説がまれにしか正しくないとすれば，それを棄却するに必要なのはサイズの大きいサンプルのみ。この場合，多くの研究で第 I 種の誤りの確率は実質的に 0。問題になるのは第 2 種の誤りのみ。

・社会科学において平均的な検定力が.5 に過ぎないことを考えると，第 2 種の誤りの確率も.5。

- Bonferroni 法のような，実験ごとの第 I 種の誤り確率を管理する方法は，検定力を.5 よりもさらに低下させてしまう(Schmidt, 1992，1996)

・特定の値のみを問題にするのであれば，非ゼロ仮説もゼロ仮説同様に確からしくない。

- しかしながら，その値を本当に検定することが現実的であるなら，非ゼロ仮説は標本から得られた結果を解釈するより現実的な基準を与える。

・研究論文の中で報告されている p 値のほとんどは，ありそうもない帰無仮説に関連したもの。

Anderson et al.(2000)：2 つの重要な環境科学誌に掲載された数百の実証研究の帰無仮説を概観

→生物学的にあり得ないような帰無仮説を多数発見(e.g. 同種の幼い個体と成体の生存確率が同じ，等)

- 社会科学領域における同様の調査を著者は知らないが，結果が違っていたら驚きだろう。

SAMPLING DISTRIBUTIONS OF TEST STATISTICS ASSUME RANDOM SAMPLING

■母集団推測モデルは，既知のサンプルからのランダムサンプリングを前提にしている。

・しかしながら，社会科学領域のサンプルは，取りやすい場所から集めてきた簡便サンプル(samples of convenience)。

・また，実験研究では，そのように局所的に利用可能なサンプルを，条件に無作為配置する。

⇒無作為配置されたケースが簡便サンプルからのものである場合，標準的な統計的検定では標準誤差がかなり大きくなる(Reichardt & Gotlib, 1999)

⇒Lunneborg(2001)：簡便サンプルから無作為配置研究の経験的標本分布を作成するためのブートストラップ法の利用方法を説明。これは 9 章で紹介する。

・NHST が理想的な形で適用できるのは，社会科学ではなく製造管理(Bakan, 1966)

STATISTICAL ASSUMPTIONS OF NHST METHODS ARE INFREQUENTLY VERIFIED

■統計的検定は通常，分布に仮定を置いており，それが崩れると p 値は正確ではなくなる。しかし，多くの研究者が，分布の仮定が満たされている証拠を示していない。…教育研究(Keselman et al, 1998)，会話・言語・ヒアリング研究(Max & Onghena, 1999)→統計研究の中での NHST と実践上のそれの乖離

・標準の F 検定のような広く使われている手法でも，仮定の侵害のせいで，正確な値が得られていることは稀(Keselman & Keselman, 1996)。

NHST BIASES THE RESEARCH LITERATURE

■雑誌の編集委員は，H_0を棄却していない研究を出版することに興味がない感がある。

・権威のある雑誌の編集委員の，H_0棄却研究を好むことに関する過去のコメント(Melton,1962)

・行動科学研究者は H_0棄却を含まない研究を投稿しない傾向にあるという調査結果(Greenwald, 1975)

・結果として，公刊された研究の大多数が H_0棄却を含んでいるということ

■統計的に有意な結果を伴う研究の見かけ上のバイアスは，以下列挙したような困難をもたらす。

①実際の第 1 種の誤りの確率は，αで示されるよりもはるかに高いかもしれない。

・効果に差がない治療法の比較研究が 100 回行われた場合，5 回は結果が有意になり，それは雑誌に発表される。一方で残り 95 の非有意な結果は公表されない。5 つの研究の第 1 種の誤りの確率は 100%。

・また，追試の失敗を研究者は報告したがらないので，第 1 種の誤りは 1 度起こると修正困難

②統計的に有意な結果がない研究を公表することへの消極性が，引き出し問題(file drawer problem)を招く。

・file drawer problem(Rosenthal, 1979)とは，雑誌や会議等で公表されない研究が存在すること。多くの“引き出し”研究が H_0棄却を含んでいないと思われる。しかし，効果が本当に 0 であれば，それら研究はゼロ仮説を棄却している公表された研究よりも，科学的に妥当。

③公表された研究は母集団効果量を過大推定する。

・小さい，あるいは中程度の効果量を研究するための大規模なサンプルがなければ，検定力の低さゆえに統計的に有意な結果を得ることは難しい。

→H_0棄却という事態は，観測効果量が母集団効果量よりも大きい時に起こる傾向にある。そしてH0を棄却した研究のみが報告されるのであれば，母集団効果量は過大推定されることになる。

e.g.Table3.2に示した6つの研究で観測された平均値差は3.58だが，統計的に有意だったものだけ(そしてこれらのみが公表されると想定して)合計すると4.22になる

TABLE 3.2
Results of Six Hypothetical Replications

Study	$M_1 - M_2$	s_1^2	s_2^2	t (38)	Reject nil hypothesis?	95% CI
1	2.50	17.50	16.50	1.91	No	–.53–6.53
2	4.00	16.00	18.00	3.07	Yes	1.36–6.64
3	2.50	14.00	17.25	2.00	No	–.03–5.03
4	4.50	13.00	16.00	3.74	Yes	2.06–6.94
5	5.00	12.50	16.50	4.15	Yes	.56–7.44
6	2.50	15.00	17.00	1.98	No	–.06–5.06
Average:	3.58				Range of overlap:	2.06–5.03

Note. For all replications, $n = 20$, $\alpha = .05$, and H_1 is nondirectional. CI = confidence interval.

NHST MAKES THE RESEARCH LITERATURE DIFFICULT TO INTERPRET

■本当に効果があるものの，検定力が50%の場合(社会科学では一般的な値)，H_0を棄却する研究は約半数で，残りはH_0を棄却しない。

・このような視点からは，研究文献は結論が出ないもののように見えてしまうかもしれない。

・問題の一端はゼロ仮説を棄却できないことが，母集団効果量が0であることを意味するという誤解に由来している。

■統計的に有意かどうかということで研究を分類するようなNHSTの結果の用い方の別の弊害：研究への不信

・私は多くの心理学の学生が「研究は何も証明しない」と言うのを聞いた。そういう学生は，研究(特にソフトな社会科学研究)で最もよくみられる言葉が「暫定的な」(tentative),「予備的な」(preliminary),そして「示唆する」(suggest)であることに気づくだろう。

・また，臨床や教育の実践家が，実践と研究が乖離しているという認識を持っていることが示されている(Williams et al.，1995，Miller, 1999)。

NHST DISCOURAGES REPLICATION

■行動科学研究者にも自然科学研究者と同じくらい追試を大事だと考えている者がいると思うのだが，社会科学領域では自然科学領域に比べて追試が軽んじられている

・Kmetz(1998)の組織科学論文13000本，経済学論文28000本の電子データベース調査

→追試と銘打っていた研究の割合はそれぞれ 0.32%と 0.18%。
→このような追試研究の比率の低さは心理学や教育の学術雑誌でも同様(e.g. Shaver & Norton, 1980)

■NHST が広く使用されていること，また先述の誤解はこの問題の一部。
・p<.01 を 100 回中 99 回結果が再現される確率と誤解している場合，追試に悩む必要を感じない。
・関連する誤信念として，統計的に有意な結果は再現されるが，帰無仮説を棄却できない結果は再現されない，というものもある。
・検定力が低い場合研究結果が一貫しないように見えることも，研究トピックへの興味を支えることにマイナスに働いているかもしれない。

■信頼区間がもっと頻繁に報告されていれば，行動科学研究での追試はもっと重要視されていただろう。
・Cohen(1994，p1002)の言葉を借りれば，行動科学のデータの信頼区間は「とても恥ずかしいくらいに大きい(so embarrassingly large)」
・信頼区間が大きいということは，研究が限られた情報しか含んでいないことを示唆。そして，これは検定結果だけが報告される場合，見えなくなるもの。

NHST OVERLY AUTOMATES THE REASONING PROCESS

■NHST の 1 つの魅力は判断過程の多くを自動化できることで，またそれは自然科学のように客観的でありたいという社会科学の要求を満たすことができる。しかし，あまりにも多くの判断仮定が自動化されてしまったという批判もある。その弊害を以下に列挙する。

①NHST の利用が二分法的思考を強める。
・NHST の究極的な結果は H_0保持か棄却の二分法。
・α=.05 のとき，研究者達は p=.06 を p=.04 と質的に異なるものとみなし，H_0に関して異なる結論に至る(実質的には同じ確率なのに)。
・Nelson(1986)の調査で，研究者の結果に対する自身は p=.05 と p=.10 のほんの少し上の地点で急激に下降することが示されている。
・NHST が二分法思考を助長することは，p が α より少しだけ大きい結果を「傾向」(trend)や「有意に近づく」(approaching significance)と記述する風変わりな習慣に寄与しているのかもしれない。しかし，p=.04 の時，「非有意に近づく」といってその結果の信憑性を割り引くことはしない。
・これらと関連する傾向として，H_0を棄却できないことを，H_1の背後の実際的仮説が正しくないことではなく，実験デザインのまずさに帰属する，というものがある。

②NSHT の使用が，データと測定プロセスから注意をそらさせてしまう。
・H_0棄却に精いっぱいになってしまうと，変数が正しく定義され測定されたかといった，データのより重要な側面を見失ってしまう
-関連する誤解：信頼性は，被調査者母集団に関するスコアではなく，テストの特性である
→このような誤解は，データに関する信頼性の報告を妨げる
-効果量の解釈には，得点の信頼性の査定が必要

③NHST 学習への多大なる時間投資が，他の方法への接触を制限する。

・NHST 以外に，さまざまな仮説やデータを扱える統計的手法があるのに，社会科学の学生は大学院においてすらも，それについてほとんど見聞きしない。プロの研究者になった人たちは，一般的にそれら方法を自分自身もしくはワークショップで学ばなければならない。

④NHST の手法は，研究知見を混ぜっ返す一方で科学的価値のほとんどない，物好きなトピックに関する研究を促進してしまうかもしれない。

・やわらかい社会科学において，研究トピックの寿命が短いことは先述した。その傾向に NHST が寄与しているという批判が存在。

-より広い理論的な説明を考慮しなければ，簡便サンプルからデータ収集し，統計的検定を実行することができる。仮に数値がランダムでも，いくつかの結果は有意になることが期待される。客観的であるかのような見た目と検定の機械的な適用が，概念的基盤の弱い研究に信憑性を与えてしまう。

NHST IS NOT OBJECTIVE AS IT SEEMS

■有意水準，H_0の様態(ゼロか非ゼロ)，対立仮説はデータ収集の前に特定されなければならないが，実践上それはまれ(これはカンニングに等しい)。より緩い基準で考えても，結果を良く見せるような変更をなしうることは，全体のプロセスを客観的というより主観的に見えるものにしている。H_0が棄却された結果だけを選択的に報告するといったことも，同様の問題を提示。

MANY NHST METHODS ARE MORE CONCERNED WITH GROUPS THAN INDIVIDUALS

■平均を分析する t や F 等の統計的検定は集団に関するもので，個人に関する情報は少ない。しかし，集団内の個人差の理解が必要な場面も存在する。第 4 章で紹介する効果量推定の方法は，このようなケース(個人)レベルの差異を分析するもの。

NHST AND SCHOOLS OF PROBABILITY

■数学，統計学，科学哲学の分野では確率に対する考え方についていくつかの流派(古典的，頻度主義，主観的等)が存在。これらの間では，確率の意味や正確な解釈について大きな隔たりが存在。そのような状況の中で，NHST は，経験的に観測可能で理論的標本分布で要約できるような，再現性のある出来事の相対的頻度として確率をとらえる立場を反映。つまり，NHST は合意のとれた確率観を代表しているわけではない。

CONTINUED USE OF NHST IS A RESULT OF INERTIA

■複数の批評家が，NHST の継続使用を，よく考えないで実行されている，空虚な儀礼的行いだとしている。

・社会科学領域の統計教育が，代替策を伝え損なっていることが，仮説検定の方法が他にないという誤信念を促進してしまっているかもしれない。

-これは事実無根。ピアジェやパブロフ，スキナー等による，心理学において影響の大きい研究は帰無

仮説を棄却せずに行われているし，自然科学は相対的に統計的検定を用いることが少ないにも関わらず，繁栄している。

・他の者は，科学において，確立された方法は一般的に変更しがたいという点を指摘する。確立された方法はパラダイムのようなものであり，パラダイムの変化は速くも簡単でもない。時に，そのような変更は世代交代を待つ必要がある。NSHT を心理学における仮説検定の標準として適用するのにも，約 20 年かかったことを思い出してみよう。

IS THERE ANYTHING RIGHT WIH NHST?

■本章でこれまで概観してきた NHST の批判は，NSHT について何か正しい部分があるのかという疑問を生じさせる。しかし，NHST の擁護者がいないわけではない。ここでは，NHST の肯定的側面を列挙する。

1.もし NHST が何もしないなら，それは標本誤差に対処する

・標本誤差は行動科学の核となる問題の 1 つ。そして，p 値は標本誤差を考慮して生成される

-そのため，複数の行動科学者は NHST を重要な要求にこたえるものと考える。このような考え方をする人は，NHST の批判者が想定するような，受動的な後追いとは異なる。

-NHST 批判者は，信頼区間がより多くの情報を伝えること，また NHST への囚われが信頼区間があまり頻繁に報告されないことの原因であることを指摘する。しかしながら，信頼区間は NHST と同様の推論の誤りのいくつかに陥りやすい。

・信頼区間は，魔法のような代替案ではない。しかし，個々の研究および追試で信頼区間を報告することは，個々の研究で統計的検定を行うことよりも科学的に信頼できる方法で標本誤差に対処するものである。

-Table3.2 の例。個々の信頼区間は標本誤差を推定しているが，それ自身も標本誤差の影響を受けている。信頼区間の重なる部分が 2.06-5.03 であるという情報は，3 つの研究で帰無仮説が棄却されたという情報より有益。

➢Schmidt(1996) : 母数に関する期待が極めて誤りであっても，研究ごとに信頼区間をプロットすることで最終的に誤りを発見することができる。

2.NHST の誤解は，方法そのものが悪いのではない。

・NHST 擁護者は蔓延した誤解を認めつつ，それは使う側の問題だと指摘する。一方，批判者は，多くの知的で高度な教育を受けた人々に誤解されるような明白な可能性を持った方法そのものに責任があると反論する。

3.より慎重に専門用語を使うことで，不必要な含意を防ぐことができるかもしれない。

・変革が示唆される一つの領域は，統計的検定の結果を報告するときの語用。

-“有意”という言葉を“統計的に”を前に置くことで制限…読者の統計的有意性と実質的有意性の区別を促進

-α より大きい小さいではなく，p 値そのものを記載する

e.g. $t(20)=2.40$ $p<.05$ でなく $t(20)=2.40$ $p=.26$ と表記

・しかし，後者にはいくつかの問題が存在。
-多くの行動科学研究で，p 値そのものが誤っている可能性について述べたが，そのような状況では，小数第 3 位とか 4 位まで正確に報告することは誤った印象を与えかねない(p 値が小さいほど結果を信用できる，というもの)。
-似たような提案は過去 50 年のうち幾度もなされたが，目に見える影響はなかったことも付け加えねばなるまい。
・NHST 批判者は，上記のような変更が社会科学での応用に関する NHST の限界を改善するものとは全く思っていない。彼らは次の言葉を的を射たものと感じるだろう
“牛フンにろうそくを立てることはできるが，そうしたところで，それがバースデーケーキにかわるわけではない。”

4.二分法的な答えを必要とするリサーチ・クエスチョンもある。
・研究を動機づけた問そのものが二分法的なものであることもある。
e.g. この介入プログラムは実行すべきか，この薬はプラセボよりも効果があるか
・また，理論が効果の方向だけを検討し，その大きさを予測しないことがある
e.g. ある読書教示はほかの教示よりも，どれくらいかはわからないが優れている
・これらいずれの場合も，効果量を推定するほうが有効。また，メタ分析により平均効果量を求めるほうが，H_0 を棄却した数を数えるよりも，結果を統合するのにより有用な方法。

5.ゼロ仮説は時に適切である。
・Frick(1995)，Robinson & Wainer (2002) 等が論じているように，ゼロ仮説の想定が適切な場合がある。複雑な研究において，1 つの独立変数が操作された時の効果を検討するケースなど。

6.NHST の技法は統計的決定(功利)論の入口である
・工学および環境学では，第Ⅰ種と第Ⅱ種の誤りの確率を，それぞれのコストで重みづけし，正味の損益を評価して不確実な状況における選択肢の合理的決定を行う。
-統計的決定論では，第Ⅰ種および第Ⅱ種の誤りの確率は恣意的に.05 や.01 に定められるわけではない。
-統計的決定論は，長期的な損害を検知できる。異なる判断のコストをドルや生存率あるいはほかの量的・客観的な測度で推定できれば，これはとても強力な方法。しかし，それは行動科学研究では可能ではない。

VARIATION ON NHST

■この節では，複数の特化した方法(基本的な NHST の変形)を確認する。それは伝統的な統計的検定の問題のいくつかを防ぐことができ，正しい状況の下では非常に有用である。

Range Null Hypothesis and Good-Enough Belts
・H_0を母集団値の範囲を示すものとして特定する方法(Serlin,1993)
・H_1も範囲仮説であるが，それは追加の分析に必要な実際的な考慮に基づいて設定される最少の結果

・仮説の値の範囲が十分域[4](good-enough belt)。この範囲で仮説の支持不支持を判断
・範囲仮説の設定は，SEM での適合度検定を除けば稀。
・また，このアプローチが実際的な違いをもたらすかには疑問も提示されている(Cortina & Dunlap,1997)

Equivalence Testing

・等質性検定(equivalence testing)は，薬理学や環境科学，生物学でよりよく知られた方法で，2 つの集団や条件を等質とみなせるかという問題に対処するもの。
・ある種の等質性検定においては，H_0 は点仮説ではなく，以下のような 2 つの範囲下位仮説に置き換えられる。各下位仮説は実際的な平均値差に対応する値の範囲を表現。下記仮説は，平均値差の絶対値が 10 以上大きければ，母集団値が等価であるとみなせないことを示している。

H_0 : $H_{01}: \mu_1 - \mu_2 < -10.00$, $H_{02}: \mu_1 - \mu_2 > 10.00$

・これと相補的な$-10.00 < \mu_1 - \mu_2 < 10.00$ は，等質性検定の十分域
・同じ α の水準で両下位仮説が棄却された場合のみ，非等質性の仮説が棄却される。同じ結論は観測平均値差の信頼区間からも導くことができる。
・ここで紹介したアプローチの第 I 種の誤りは，グループが等質ではないのに等質であると言ってしまう確率。これは消費者にとってのリスクである。
・もし，第 1 種の誤りの確率が消費者でなく生産者のためのものであれば，適切な帰無仮説は

H_0 : $-10.00 \leqq \mu_1 - \mu_2 \leqq 10.00$

となるが，これは観測平均の信頼区間の下限が 10 を超えるか，上限が-10 を下回るかのいずれかの場合に棄却される。

Inferential Confidence Intervals

・Tyron(2001)の統計的差異，等質性，非決定性(inditerminancy：統計的に有意でもなければ等質でもない)に関して平均を検定するための統合的アプローチ
-推測信頼区間(inferential confidence intervals)に基づく
…個々の平均値に算出される加工された信頼区間
…推測信頼区間
= {(平均値の標準誤差)×(修正要因で調節された両側の境界 t 値)} /個々の標準誤差の和
…修正要因は平均値差の標準誤差の比に等しく，その値は.70～1.00
→推測信頼区間は，同じ平均に関する標準の信頼区間より概して狭い
・このアプローチで統計的差異が見出されるのは，2 つの推測信頼区間が重ならない時。
-この差異にかかわる確率は，ゼロ仮説と方向性のない対立仮説についての t 検定の場合と同じ。
…つまり，差の検定に関しては通常の NHST と同じ結論に至る。
・統計的な等質性があると結論されるのは，2 つの平均値の最大可能差異(maximum probable difference)が，等質性仮説で取るに足りないと考えられる量よりも小さい時。
-最大可能差異とは，2 つの推測信頼区間の上限の最大値と下限の最小値の差

[4]定訳ではないです。

e.g.　10.00-14.00 と 12.00-18.00 の 2 つの推測信頼区間→最大可能差異 18.00-10.00=8.00
→この値が等質性仮説で設定した範囲に収まれば，統計的に等質であると考える。

・統計的差異にも等質性にも達しない平均の対比は，統計的非決定であり，いずれの仮説にも反しない。
・Tyron は，以下の 3 つの理由からこの方法がより誤解を招きにくいと主張している
(a) 帰無仮説が顕在的ではなく潜在的
(b) 差異と等質性の両方をカバーしている
(c) 統計的非決定という 3 つ目の結果が利用可能なことで，有意水準に届きそうな結果の“傾向”という解釈を減らすことができる。
…このアプローチが肯定的な結果をもたらすかどうかは，現時点では不明。

Three-Valued Logic

・Kaiser (1960)は NSHT 二分法論理を三元評価の論理に置き換えることを最初に提唱した社会科学者。
・三元評価では，方向の違う片側ずつの対立仮説を認める。そして，予測と方向が異なっていれば，統計的に有意な結果でも実質的仮説に反するものと考えることを許容。
・Harris(1997b)は極めて明晰で現代的な三元評価論理の説明を行っているが，この論理はあまり使われていない。

WHAT DO WE DO NOW? BUILDING A BETTER FUTURE

■統計的検定の批判を考慮した後，我々は以下の行為の道筋の 1 つを選ぶことができる。
①何もしない：つまり，過去 50 年そうしてきたように統計的検定を使い続ける
②統計的検定の使用を完全に止める：大学で教えるのをやめる。それらを使った研究を出版するのを拒絶することで効果的にその使用を禁止する。この選択肢は仮想的で急進的にすら聞こえるだろうが，何人かの高名な研究者がこのような禁止を求めている（e.g. Hunter, 1997;Schmidt, 1996）
③上述の両極端の間を行く道を探す：研究文脈によって，異なる程度（全くなしからある程度重要なものとして）での統計的検定の使用を求める。ただし，その使用には厳しい必要条件を設ける。
・選択肢①は許容できない。行動科学研究の進展に否定的な含意があるから。また，短期的には心理学の雑誌における統計的検定の禁止はありそうにもない。ゆえに，最初の 2 つの選択肢は除外される。

■以下の勧告は 3 つ目の選択肢に基づくもの。その主題は，「統計的検定は人文科学との差別化，また主要な役割モデルたる自然科学への接近に苦労した思春期を通じて，心理学や関連行動科学者を助けた。しかしながら，行動科学はその思春期を脱し，物事をなす新しい方法へ成長せねばならない」
【勧告】
①効果が存在するかが未知であるような極めて探索的な研究が，NHST が主要な役割を果たすことが適切であるような唯一の場である。
②もし統計的検定が使用されたら，(a) 検定力に関する情報が報告されねばならない，(b) 帰無仮説が現実的なものでなければならない
③あらゆる種類の行動科学研究において，NHST の結果のみから研究結果を記述することはもはや許容できない

④“有意（significant)”という言葉を我々のデータ解析の言葉から除外せよ。それを日常的な意味で何某かの本当に注目すべきことあるいは重要なことを指す場合にのみ使用せよ。

⑤可能な場合はいつでも，主要な結果の効果量の推定値と信頼区間を報告し解釈することは研究者の責務である。これは，H_0棄却についてのみ効果量を報告するという意味ではない。

⑥実質的(理論的，臨床的，実践的)有意性を示すこともまた研究者の責務である。統計的検定はこの目的のためには不十分である。

⑦再現は標本誤差への最善の対処法である。

⑧統計手法の教育は改革される必要がある。NHST の役割に今よりもはるかに重きを置かないようにし，より多くの時間を結果が実際的な有意性を持つのかの確認する方法，結果を追試する方法を学生に示すことができるようにすべき。

⑨研究者は，効果量と信頼区間の計算について，統計的ソフトウェアからのさらなる助力を必要とする。

A Primary Role for NHST May Be Suitable Only in Very Exploratory Research

・関係が標本誤差の期待水準よりも大きいか否かという二分法的問いに応える NHST の能力は，いくつかの新しい研究領域では有用。

・ただ，NHST の限界を考えれば，有用なのはごく短い期間。効果があるという証拠が存在すれば，次の段階ではその大きさの推定や実質的有意性の評価が次のステップ。より進んだ効果研究は，個々の仮説ではなくモデルを検定する技法，特にモデル適合の手法(SEM,HLM,潜在クラスモデルなど)を必要とするだろう。

Report Power for Any Use of Statistical Tests, and Test Only Plausible Null Hypothesis

・報告されるのは観測検定力ではなく，事前検定力(a priori Power)。

・多くの結果が H_0を棄却できないとき，検定力の報告は特に重要

-読者が検定力の低さゆえに 否定的な結果が生じているのか否か判別できるはずだから

・結果が否定的な場合に，研究文献で検定力を報告している例はほとんど見ないだろう。これは出版バイアスのため。しかし，より偏っていな研究では，検定力の報告はより重要な意味を持つ。

-結果の頻度の小ささを誇張するような小さな p 値は現実的でない帰無仮説の下で生じる。

It Is Not Acceptable to Describe Results Only on the Basis of NHST Outcomes

・ここまで考慮した NHST の欠点がこの勧告に根拠を与える。

・また，雑誌編集委員や差読者にとって，NHST の結果が採択か棄却を決める重要な点となってはならない。

Stop Using the Word “Significant” in Connection With NHST

・$p < \alpha$を指し示す言葉に「有意」という言葉を選んだのは非常にまずかった。統計家はそれに大きいとか重要とかいう含意がないことを理解しているが，非統計家にはわからない。

→我々行動科学研究者は，通常の語用（重要さ，有意味性，実質性を示す）で使うべき。

・H_0が棄却された場合には「統計的」という言葉を使うだけで十分。

H_0：ρ=0　の棄却→「統計的な関連が存在する証拠」

H_0：μ_1 = μ_2 の棄却→「統計的な平均値差の証拠」

*統計的な効果が有意味な効果である場合もあるが，それは NHST の預かり知るところではない。

・上記の単純な語用は，「統計的に信頼できる(statistically reliable)」よりも好ましい。

→「信頼できる」は再現可能性を含意するが，p 値はその情報を全く含まない。

・少なくとも，有意という言葉を使うのであれば「統計的」という前置きで意味を制限すべき(Thompson, 1996)

Whenever Possible, Researchers Should Be Obliged to Report and Interpret Effect Sizes and Confidence Intervals

・だんだんと多くの雑誌が効果量の報告を求めるようになっているが，これはこの勧告を支持すること。

・効果量推定値の信頼区間を算出することはさらに好ましい

-区間の幅が，観測効果量にかかわる標本誤差の量を直接に示す

-観測効果量を与えるような母集団効果量の範囲を算出することが可能になる

・いくつかの複雑なデザインの下で効果量を算出すること，ある種の統計値から信頼区間を算出することは必ずしも可能でない場合もあるが，たいていの行動研究では可能。

Researchers Should Also Be Obliged to Demonstrate the Substantive Significance of Their Results

・帰無仮説の棄却は実質的有意性は意味しない。それゆえ，研究者は研究の面白さや重要性を説明する他の枠組みや参照点が必要。

・効果量の算出がその第一歩。もしあるなら，領域におけるメタ分析の結果を参照する。

Replication is the Best Way to Deal With Sampling Error

・雑誌や助成団体が研究者に結果の再現性を要求すれば，それは強い声明となる。

-研究者への要求が厳しくなり，公刊される研究は少なくなるだろう。

-しかし，公刊される研究の質は改善するかもしれない。

・また，再現性の要求は，一時的に流行してすぐ廃れる研究テーマのいくつかをふるい分けるだろう。

・このような要求は，領域に大きな影響を与える独創的な研究に対してはいくらか緩和されるべきだが，予期しないような驚くべき結果についてはより厳格に求められるべき。

Education in Statistical Methods Should Be Much Less NHST-Centric and More Concerned With Replication and Determining Substantive Significance

・多くの初級統計の授業で，NHST の技法は頂点にあるかのように示される。

-統計の学部レベルテキストが NHST を基礎に内容を提示。大学院でも新たな NHST 技法とその使用法しか教えない。また，多くのテキストは古典的統計的検定以上の方法―例えば効果量―を重視しない。

・いくつかの初級の講義ですでに教えられているようなトピックは，もっと注目されるべき。

-効果量指標は，相関，標準偏差の割合，ある値に落ち着く得点の割合に過ぎない。これらは初級の講

義で取り扱われている基礎的な統計指標。

-しかし，これら指標の，古典的記述統計と推測統計以外への応用は教えられていない。

・NHST を重視しないことにより，学生はより単純な分析の方法を選ぶようになるだろう。そして，このことは，有意味あるいは注目すべき効果(それらは比較的単純な統計指標や図表等を見れば明白)を検出するのに統計的検定が必ずしも必要でない，ということを正確に認識する助けとなるだろう。

・また，データにより近いレベルで結果を記述することは，学生のコミュニケーションスキルの発達を促進するかもしれない。

-このことは，政策決定者に自分たちの研究結果の含意を説明する必要を伴うキャリアを歩む学生にとって特に重要(McCartney & Rosenthal, 2000)

・また，統計と研究法の講義を統合する必要がある。現在は両者が別々の講義で教えられていて，学生は実際の使用に関する感覚を得ないままにデータ分析を学んでいる。

Statistical Software Should Be Better at Computing Effect Sizes and Confidence Intervals

■ほとんどの一般統計ソフトウェアは極めて NHST 中心。

・だんだんと，それらソフトウェアが，少なくとも数種類の効果量指標をオプションとして出力するようになっているということは励みとなるもの(今までの議論からすれば NHST のほうをオプションにすべきだけども)。

・多くの効果量指標が利用可能であり，少なくともより広く使用されているような指標は各分析のコンピュータープログラムの出力で利用可能であるべき。

-所与の統計的分析について，複数の異なる効果量指標が利用可能であることも事実。

■多くの現代的一般統計ソフトウェアは母集団平均や回帰係数についての信頼区間をオプションで出力。しかし，母集団効果量の信頼区間も出力すべき。

CONCLUSION

・統計的検定は，社会科学にとってロールシャッハのインクブロットテスト(集合的な)のようなもの。

- 我々がそこに見出すものは，事実よりも欲求充足により関連が強いものであった。
- この不思議な集合思考が集積的科学としての心理学と関連領域の発展を阻害してきた。

・また，行動科学研究の特徴と，統計的検定の結果が正確であるための要件に不整合が存在する

- 現実性のある帰無仮説，ランダムサンプル，分布の仮定，推定検定力，p 値の正確な意味…行動科学ではこれらの**いずれも**一般的に保持されていない。

・本章では，古典的統計的検定の役割を縮小するようないくつかの提案を行った。

- 関心のある結果全てについて効果量と信頼区間を算出(帰無仮説が棄却されたものだけでなく)，結果の実質的有意性の評価(統計的有意性だけでなく)。
- 追試こそが最も重要な変革。