Jones-Tukey定式化註釈

—「10%有意水準」を正当化する新しい統計的な考え方について—

月　元　　敬（岐阜大学）

A commentary on the Jones-Tukey formulations:
On a new statistical view of justifying the 10% significance level

Takashi TSUKIMOTO (*Gifu University*)

近年，心理学を含む様々な科学コミュニティにおいて，*p*値だけに頼る統計の使用を改めようとする「統計改革」が始まっている。この動きは，サンプルサイズを増やせば，操作の効果が小さくても有意となってしまうという帰無仮説検定に対する批判が大元になっている。本論文の目的は，現時点では心理統計学の専門家によって呼びかけられている「統計改革」に含まれておらず，それ故，心理統計学の非専門家にはほぼ知られていないであろう，検定に対する非常に優れた新しい考え方であるJones-Tukeyアプローチに関して註釈することである。

Key words：Jones-Tukey approach, significance testing, reversal errors

心理学者が「結果が出た／出なかった」と言う場合，それは「統計的に有意となった／ならなかった」という意味である。「有意でなかった」ということも本来は「結果」であるのだから，この特殊な言い回しは実証研究の従事者が「統計的に有意である」ことを重視する（あるいは崇拝する）姿勢を示すものであると言える。

近年は，有意であることだけでなく信頼区間と効果量を報告する流れになっている（e.g., 大久保・岡田, 2012; Wilkinson & APA Task Force on Statistical Inference, 1999）。最も極端な場合，例えば，「検定など大惨事（disaster; 他にも大失敗という意味もある）だから捨て去ってしまおう」という主張がなされたり（e.g., Hunter, 1997），2015年に学術誌*Basic and Applied Social Psychology*は統計的検定及びそれに類する統計学的処理を禁止する立場を宣言したり（Trafimow & Marks, 2015）といった「全面廃止」の動きも出てきている。また，検定力分析（e.g., 南風原, 2014; 豊田, 2009）やベイズ統計学（e.g., 豊田, 2015, 2016）の必要性が訴えられるようになってきたのも「統計改革」が日本の心理学界で始まっていることを物語っている。

とは言え，帰無仮説検定は，有意水準という明確な基準があることが強みである。つまり，有意水準は慣習で定められている確率（慣例的に５％）であるものの，研究者にある程度の客観的な認識法，すなわち，有意か否かという2値的な判断を可能にしてくれる基準である。特に，他の科学領域とは異なり，尺度の精度があまり高いとは言えない心理学においては，効果の大きさよりも効果の有無が専ら心理学者の関心事とならざるを得ないこともあり，それ故，帰無仮説検定は批判を浴びる存在でありながらも今後も「消滅」することはないだろう。

以上のように，統計改革の真っ只中でありながらも心理学者は自身の実証的研究で「結果が出る」すなわち「有意である」ことを絶えず期待する。そのために心理学者はデータの信頼性や妥当性を高めるような計画や測定を考えるのである。言い換えれば，彼らの工夫は全て，条件間の差があるのであればそれを正しく検出できる確率，いわゆる「検定力」を上げようとす

る試みの一環ということになる。

一般に，検定力を高める要因として次の4つがある。(a) サンプルサイズが大きい方が良い。(b) 検出しようとする真の差が大きい方が良い。(c) 分散が小さい方が良い。(d) 有意水準αが大きい方が良い。これらのうち研究者の努力の範囲であるのは (a) と (b) である。カウンタバランス等によって剰余変数の統制を行うのは，あり得る差が大きいのであれば，できる限りそれに近い差をサンプルから見出そうとする研究者各人に許された手立てである。しかし，条件を隔てる操作による効果の現れ方にかなりの個人差があるかもしれず，その意味で (c) は統制の施しようがない。

では，(d) の有意水準αの設定はどうか。「有意水準５％」というのはあくまでも慣習的なものであり論理的な根拠があって設定された確率ではない（歴史的に見れば，有意水準を５％にする慣習は「Fisherの個人的好み」から始まったとも言える)。その意味で，基本的には研究分野や領域で「研究者の判断によって決められる」はずのものである。しかし，どんな世界でも同じように慣習が固定化や呪縛と紙一重であり，現在の心理学的研究において「有意水準の設定」に自由など存在しないように感じられるのは私だけではないだろう。

そんな中，Jones & Tukey（2000）は「有意水準10%の（両側）検定を行えば良い」と提唱した。これは現在の実証研究でもしばしば見られるような，*p*値が10%未満である時に「有意傾向である」ことを積極的に言って良いということではない。

「有意水準を10%にする」というJones & Tukey（2000）の主張は，ただ検定力を高められるからというよりもむしろ仮説検定の論理的な意味を明確にしたという点で，現在主流となっているNeyman-Pearsonアプローチよりも優れていると（少なくとも私には）思われる。Jones & Tukeyが論文を*Psychological Methods*という心理学系の学術誌に投稿したのは，前述したように，効果の大きさよりも効果の有無に関心が強いからこそ，Jones-Tukeyアプローチが心理学的研究には非常に意味があると見込んだからではないだろうか（定かではないが)。それにも関わらず，「統計改革」の１つとしてJones-Tukeyアプローチを紹介しているものは見当たらない。その理由の１つは，やはり「習慣」に由来すると思われる。

本稿の目的は，非常に理に適っているにも関わらず統計改革として議論されることのないJones-Tukeyアプローチについて，その基本的な考え方を註釈し，仮説検定の意味について再考する機会を提示することである。

Jones-Tukey定式化

話を単純にするため，２つの平均値の差に関する*t*検定を例とする。従来的な考え方では，２つの平均値μ_Aとμ_Bに関する結論候補は，(a) $\mu_A > \mu_B$, (b) $\mu_A < \mu_B$, (c) $\mu_A = \mu_B$の３通りある。(c) は伝統的な仮説検定では帰無仮説として設定され，入手データの有意性が判断される。非有意の場合，Fisherアプローチでは帰無仮説を「保留」するのに対し，Neyman-Pearsonアプローチでは帰無仮説を「採択」する。この違いのため，Neyman-PearsonアプローチではタイプⅠエラーの危険を小さく抑えつつ，同時にタイプⅡエラーも同様に小さく抑えなければならないという検定力の発想につながるのである(FisherアプローチはタイプⅠエラーのみを考慮)。

帰無仮説 $\mu_A = \mu_B$は「極めて大胆な仮説」と言われる（山内, 2008)。それは帰無仮説が「虚偽命題」だからである。$\mu_A = \mu_B$を額面通り受け取れば，小数点以下の値がどこまでも等しいことを意味する。しかし，そのようなことはあり得ず，必ず $\mu_A \neq \mu_B$ なのである（Loftus, 1996; Tukey, 1991)。すなわち，

$$P(\mu_A = \mu_B) = 0$$

$$P(\mu_A \neq \mu_B) = 1$$

が必ず成立する。したがって，論理的には，Neyman-Pearsonアプローチにおける非有意に対する捉え方は水泡に帰すことになる（はずが，そうならないほど甚だしく慣習化してしまっているのが現状である)。以上のことから，２つの平均値μ_Aとμ_Bに関する世界のありようは間違いなく (a) $\mu_A > \mu_B$, (b) $\mu_A < \mu_B$ のいずれかで

ある。

となると，心理学者の関心事である効果の有無よりもむしろ，(a) か (b) かという「差の方向」を問う方が遥かに重要である。つまり，「差の検定」であった t 検定は「差の方向の検定」へと大きく変貌することになる。さらに，t 検定の結果，非有意であった場合，「差の方向不定（決定の証拠として不十分）」というFisherアプローチに類する結論とせざるを得ない。以上から，t 検定による結論として，(a) $\mu_A > \mu_B$，(b) $\mu_A < \mu_B$，(c) 方向不定のいずれかを受け容れることになる。

以上がJones & Tukey (2000) による提案の基本である。ここで，t検定の流れとしてまとめておく。サンプルサイズに沿う自由度の t 分布において検定統計量 t が，

① もし上側の棄却域に入れば，$\mu_A < \mu_B$ を棄却し，$\mu_A > \mu_B$ と結論する。
② もし下側の棄却域に入れば，$\mu_A > \mu_B$ を棄却し，$\mu_A < \mu_B$ と結論する。
③ いずれも棄却できない時，方向不定（証拠不十分）とする。

リバーサルエラー（reversal errors）

Jones-Tukeyアプローチは帰無仮説が正しい確率はゼロであることを初めから認める立場であるので，従来的なNeyman-PearsonアプローチにおけるタイプⅠ／Ⅱエラーは起きようがない。しかし，先の定式化によって，タイプⅠ／Ⅱエラーとは性質の異なるエラーが浮上する。すなわち，結論に伴うエラーは「差の有無」ではなく「差の方向」にしか起きない。Jones & Tukey（2000）はそれを「リバーサルエラー」と呼んでいる。

表1はJones-Tukeyアプローチにおけるエラーを表している。なお，Jones & Tukey（2000）は（不親切にも）この表を掲載していないので，タイプⅠ／Ⅱエラーを説明するために用いられる表を参考に作成した。この表で注意が必要なのは，結論が方向不定である場合であるが，これについては後述する。

リバーサルエラーについて，t検定（有意水準5％での両側検定）に当てはめて考えてみよう。t統計量が上側か下側のいずれかの棄却域に入り，それに対応する結論を出したが，それがリバーサルエラーであったとすると，その確率 P(RE) は2.5%になる。少しややこしい表現をすれば，有意水準を5％とした時，リバーサルエラーを犯す確率（危険率）は2.5%になるということである。

より一般的には，有意水準 α に対し，

$$P(\mathrm{RE}) = \frac{\alpha}{2} \times P(\mu_A > \mu_B) + \frac{\alpha}{2} \times P(\mu_A < \mu_B)$$

$$= \frac{\alpha}{2} \times \{P(\mu_A > \mu_B) + P(\mu_A < \mu_B)\}$$

$$= \frac{\alpha}{2} \times P(\mu_A \neq \mu_B)$$

$P(\mu_A \neq \mu_B) = 1$ であるから結局，

$$P(\mathrm{RE}) = \frac{\alpha}{2}$$

となる。つまり，リバーサルエラーを犯す確率は設定した有意水準 α のちょうど半分になる。このことから，リバーサルエラーの確率を5％に抑えるのであれば，$\alpha = .10$ の両側検定を行えば良いことになる。これが，Jones & Tukey

表1　Jones-Tukeyアプローチにおけるエラー

		真実(母集団の状態)	
		$\mu_A > \mu_B$	$\mu_A < \mu_B$
検定の結論	$\mu_A > \mu_B$	正	リバーサルエラー
	$\mu_A < \mu_B$	リバーサルエラー	正
	方向不定	チャンスの無効化	チャンスの無効化

（2000）が提案した10%有意水準に正当性を与える新たな考え方である。

以上の考え方には不可思議な理屈は何もないように思われる。ただ，疑問を持たれる可能性もないわけではないだろう。例えば，Jones-Tukeyアプローチは，

$$P(\mu_A = \mu_B) = 0$$

を完全に認めてしまっているので，検定の結果，非有意である場合には，Neyman-Pearsonアプローチのように「帰無仮説を採択する」すなわち「$\mu_A = \mu_B$」という結論を出すことは論理整合性の観点から不可能である。にも関わらず，たとえ形式的であれ，帰無仮説分布を用いることに論理的な問題はないのだろうか。

Jones & Tukey（2000）が直接この疑問に回答しているわけではないが，結論を言えば，帰無仮説分布を用いることに論理的な問題はないと思われる。鍵となる考え方は「報告すべきp値」に関係している。

Jones & Tukey（2000）は，報告すべきリバーサルエラーの確率pは両側確率ではなく，片側確率とするのが適切であると述べている。検定統計量tが十分に大きければ $p \cong 0$ であり（このことは他の検定統計量にも当てはまる），t値がゼロに近い時は $p \cong .5$ である。つまり，$0 < p \leq .5$ である。$p = .5$ の時は「母平均の大小関係は五分五分である」ことを意味するので，いずれかの方向を結論すれば，それは文字通り「当てずっぽう」であり，それが正しい確率はチャンスレベルに過ぎない。これは，コイン投げで「表」に賭けると，当たるのも外れるのも50%であるという不確定な状態であることと同じである。Jones-Tukeyアプローチにおける検定の機能が「差の方向」を判断することであることを踏まえれば，一方の方向に肩入れしない，中立な立場から方向を見極めるためには，帰無仮説分布を用いることが最も自然なのである。これは，ベイズ統計学における「理由不十分の原則」という考え方に類似していると言えるだろう。

チャンスの無効化

Jones & Tukey（2000）は，差の方向性を判断できないという結論を下すことを「チャンスの無効化（wasted opportunity）」と呼び，これはエラーではないと考えている。確かに，リバーサルエラーに比べれば，チャンスの無効化は「$\mu_A = \mu_B$」と結論を主張しない点でエラーの重みは明らかに低い。Jones & Tukeyはこのような性質を踏まえて，これをエラーとは考えなかったのかもしれない。

しかし，Jones & Tukey（2000, p.412）は，「チャンスの無効化を最小にしつつ，リバーサルエラー率をコントロールする必要がある」と述べている。必ず $\mu_A \neq \mu_B$ であるならば，存在する差（の方向）を結論づけられないという「チャンスの無効化」は，形式的にタイプⅡエラーに対応させることができる。したがって，検定力に関する理論をJones-Tukeyアプローチに形式的にかつ論理整合的に導入することができることは明らかである。

まとめ

以上の議論で明らかになったように，Jones-Tukeyアプローチでは帰無仮説やエラーに関する考え方が従来のものと異なっているが，形式的には有意水準を10%に設定することと同じである。つまり，この新しい考え方では，リバーサルエラーが起きる確率を５%に抑えようとするならば有意水準を10%にしても良いということが論理的に正当化される。そして，有意水準を2倍に設定できることから，必然的に検定力も上昇することになる。

但し，従来的なアプローチにおいて有意水準を５%とすることに必然的な根拠がないのと同じく，Jones-Tukeyアプローチは有意水準を10%とする必然的な根拠を与えるものではないことに注意しなければならない。別の言い方をすれば，Jones-Tukeyアプローチは「慣習と検定の再定式化」に関する提案なのである。

Jones-Tukeyアプローチは，従来的な考え方よりも検定の意味をより明確にしている点で，

非常に優れた考え方であると思われる。もちろん，Jones-Tukeyアプローチにおいても，サンプルサイズが増えれば有意な結果になりやすいという問題は依然として残る。しかし，単に有意水準を上げて検定力を高めるというのではなく，その背後の考え方は「統計改革」の中に含めてもおかしくないほどであるというのは言い過ぎであろうか。

このアプローチが統計学の専門誌ではなく*Psychological Methods*に投稿され，そして掲載されたことは重大な意味があるように（少なくとも私には）思われる。Jones & Tukey (2000) は，「統計的検定なしに心理学研究はあり得ない」というパラダイムに従う心理学者に「検定に対する再考を迫る狙い（願い）」があったのではないだろうか。

この論文が発表されてから20年弱が経過した今，「統計改革」に関わる書籍において，Jones & Tukey (2000) の論文を引用している「統計改革」関連の書籍は私の知る限り存在しない (e.g., Cumming, 2012; Kline, 2013; 大久保・岡田, 2012)。検定の内容を信頼区間が包含しているのだとすると，検定に対する考え方が異なれば信頼区間が意味することも異なってくるはずである。また，サンプルサイズ，有意水準，検定力，効果量の4つは，他の3つが決まると残りの1つが決まるという関係であるから，有意水準の意味が変われば，他の3つも少なからず意味解釈の変化が起きることになる。Jones-Tukeyアプローチは「統計改革」の重要な概念に対して無関係ではないのである。もちろん，本稿はJones-Tukeyアプローチのみの採用で「統計改革」が可能であると主張しているのではない。ただ，残念ながら，*p*値の問題とベイズ統計学とがかなり声高に叫ばれてきているという趨勢により，Jones-Tukeyアプローチに基づく「統計的検定の概念的転回」が議論されることはほぼないだろうという悲観的な見方にならざるを得ない。

引用文献

American Psychological Association (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

南風原 朝和 (2014). 続・心理統計学の基礎―統合的理解を広げ深める―　有斐閣アルマ

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3-7.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*, 411-414.

Kline, R. B. (2013). *Beyond significance testing: Reforming data analysis methods in behavioral research* (2nd ed.). Washington, DC: American Psychological Association.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161-171.

大久保 街亜・岡田 謙介 (2012). 伝えるための心理統計―効果量・信頼区間・検定力―　勁草書房

豊田 秀樹 (2009). 検定力分析入門―Rで学ぶ最新データ解析―　東京図書

豊田 秀樹 (編著) (2015). 基礎からのベイズ統計学―ハミルトニアンモンテカルロ法による実践的入門―　朝倉書店

豊田 秀樹 (2016). はじめての統計データ分析―ベイズ的＜ポスト*p*値時代＞の統計学―　朝倉書店

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1-2.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100-116.

Wilkinson, L., & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

山内 光哉 (2008). 心理・教育のための分散分析と多重比較―エクセル・SPSS解説付き　サイエンス社