生成量と研究仮説が正しい確率
–ポストp値時代の統計学–

早稲田大学文学学術院 豊田秀樹

有意性検定は、手続きが客観的で、結果が白黒はっきりつくので明快である。p 値が 5%を下回れば、統計的に有意な差があると認められ、その判定自体には異論がでない。投稿者側は安心して論文が書ける。投稿者側ばかりでなく査読をする側も「有意なら差がある、そうでなければ差が示されていない。」と自動的に判断でき、単純明快である。ボランティアであることが多い査読において、自動的・機械的に判断可能であるという有意性検定というツールは、短時間で採否を判定できるから、正直なところ大変ありがたい。しかし客観的な手続きとその目的とがズレていると、それがほんのわずかなズレであっても、ズレの領域に論文が押し寄せる。

有意性検定による客観的な手続きと、その目的とのズレとは p 値による有意差と学術的に意味のある差とのズレである。たとえば実験群と対照群の平均の差で考えてみよう。有意性検定における典型的な帰無仮説は、母平均の差に関する

$$\mu_{\text{実験群}} - \mu_{\text{対照群}} = 0 \tag{1}$$

であり、これを棄却することによって統計的な有意差を示す。しかし固有技術的には差が 0 以上なだけでは意味がない。たとえば記憶術の提案研究で 100 点満点の試験で 0.1 点の母平均差は、(1) 式の命題では偽であるが意味がない。同様にダイエット法の提案で 1 か月頑張って 3g の減量の有意差を示しても意味がない。帰無仮説 (1) 式は、データ数が大きくなるにつれて、検定力が大きくなり、いずれ棄却され、有意差が見出される。価値のある研究をするより、価値はなくてもデータを少しばかり余分にとって有意差のある研究をするほうが、ずっとずっと楽である。研究者であれば誰しも論文を掲載したいという切なる願いを持っている。正当なルールの範囲内なのであるから、当然、このズレに研究者と論文は悪意なく殺到する。これが公刊直後から全く影響を与えない論文が大量に公刊される最大の原因となる。

この問題を解決する方法として、ベイズ的アプローチによる研究仮説が正しい確率の利用がある。たとえば実験群と対照群の母平均の差の関数である生成量が基準点 c より大きい確率

$$p(\text{生成量} > c) \tag{2}$$

を計算することが可能である。生成量としては、平均値の差・効果量・非重複度・優越率・閾上率などがある。実質科学的に意味のある差が存在する確率そのものを評価することによって、手続きとその目的とのズレを解消する。ただしこの方法では、投稿者は意味のある基準点 c を主張しなくてはいけないし、査読者は評価しなければなくなる。たとえば記憶術の提案研究ならば、10 点上昇、偏差値 5 アップなどの研究目標である。ただしデータで示された効果が仮に小さくても

$$c = f(\text{実験を規定する条件}) \tag{3}$$

のような実質科学的説明が与えられており、実験条件を変更すれば厳しめの基準 c をクリアできることを（社会に貢献し得ることを）、査読者に納得させられるならば、基礎研究の論文として採択すべきである。

有意性検定のような、たとえば 5 %というような紋切り型の基準がないと査読側の手間は大変になる。それでもなお、ドメイン知識を有するレフリーが、実感の伴った研究仮説が正しい確率や基準点を、ひとつひとつ丁寧に評価し、論文の採否を決定すべきである。遠回りのように見えて、それこそが社会に役に立つ論文を増やす近道である。p 値に限らず、数理統計学にだけ依拠する指標に基づいて自動的・機械的に査読の判定を行う方針は改めるべきではないだろうか。