医療職のための統計シリーズ

医療職のための学び直し－研究デザインから論文報告までの生物統計学の道標－
第8回　検定とp値

坂巻（サカマキ）　顕太郎（ケンタロウ）*

Ⅰ　はじめに

介入効果を評価する研究に限らず，論文に解析結果を載せる際，「統計的に有意（statistically significant）」「$p<0.05$」などと検定から得られた結果の一部のみを載せていることがある。また，p値（Ⅱ（5）節）が0.051であった解析結果について，“marginally significant”などといった表現を記載している論文も散見される[1]。p値の記載については，R，Python，SAS，SPSSなどのソフトウェアを使って得られた結果をそのまま転記していることがあり，「p-value <2.2e−16」「$p=0.000$」などと記載されていることがある[2]。これらはすべて，誤った検定結果の記載と考えられている。

論文にどのような解析結果を記載すべきかはいくつかの考え方があり，ジャーナルごとに異なる。例えば，*Epidemiology*は，検定（標準的ではあるが，恣意的なカットオフ値をp値が超えたかどうかという判断）の結果やp値の記載を勧めていない[3]。一方で，*New England Journal of Medicine*は，p値はいまだに重要な役割を持っていると考えており，医師がどの治療を使用するか，規制当局がどの治療を認可するか，といった意思決定との関連から，丁寧にデザインされた研究（well-designed randomized or observational study）において，研究開始前に設定した主たる仮説（検証すべき仮説）に対して，研究開始前に設定した解析方法を用いて得られたp値を論文に記載することは容認している[4]。

検定の利用や記載に関して様々な考え方がある理由の一つに，検定の誤用がある。議論の内容についてはWasserstein and Lazar[5]やその和訳[5]などを参考にしてほしいが，簡単には，p値に対する誤解，p値の誤用が問題といえる。検定を利用する際は，特に，「科学的な結論や，ビジネス，政策における決定は，p値がある値を超えたかどうかにのみ基づくべきではない」「p値や統計的有意性は，効果の大きさや結果の重要性を意味しない」[5]という二つの点に注意してほしい。例えば，新たな降圧薬（試験治療）が既存の降圧薬（標準治療）に対して効果がないかどうか（無効でないかどうか）の判断に検定を用いる場合，具体的な効果の大きさに関する情報は（単独な）検定からは得られない。そのため，仮に「統計的に有意」であっても「臨床的に有意（clinically significant）」とまでは判断できないということになる。介入効果を評価する研究では，効果の点推定値や95%信頼区間（**連載第7回参照**）を論文に記載すべきであり，必要があれば，意思決定に関する情報として，検定を適切に利用し，その結果を論文に記載するという方針が望ましい。

検定を適切に利用するには，検定が何かを理解する必要がある。検定にはいくつかの種類があり，有意性検定（significance test），統計的仮説検定（statistical hypothesis test），ベイズ流（Bayesian）検定などがある[6]。これらの違いは，「確率」の定義や「検定手順」などであるが，共通するところもあり，統計家でも正確に理解するのは難しい。統計学の入門的なテキストでは，統計的仮説検定（Ⅲ節）をもとに検定を説明しているものがあるが，実際に使われている多くの検定とp値は，有意性検定と統計的仮説検定を混ぜたものであるため，非統計家が適切に検定を理解することが困難となっている。実際，アメリカ統計協会（American Statistical Association）のp値に関する声明における説明[5]は，有意性検定によってはいるが，統計的仮説検定を含んでいるようにみえる。

*横浜市立大学データサイエンス推進センター特任准教授