p 値と仮説検定：どう教えればよいか

岡山大学　環境生命科学研究科（環）　坂本　亘

１．はじめに

R などの統計ソフトウェアで検定のプログラムを実行すると p 値が出力され，利用者は p 値を有意水準と比較して仮説を棄却するか否かの判断を下す．一方，多くの統計学のテキストでは，意思決定の手段としての仮説検定の理論や手順の説明に主眼がおかれ，p 値は検定統計量の分布に絡めてやや短めに説明されている．このようなソフトウェアとテキストの間に生じている相違が，特に初学者にとって，統計的検定に対する誤解の一因になっているのではなかろうか．

２．Fisher の有意性検定と Neyman-Pearson の仮説検定

Fisher (1925) はある仮説からの乖離の程度を，その仮説のもとでの検定統計量の分布の裾確率 P によって表した．標本に基づく P の値が十分に（例えば 0.05 よりも）小さいとき，その乖離は有意であるとした（仮説の棄却／受容には言及しない）．これが p 値の由来である．

Neyman & Pearson (1928, 1933) は仮説が正しいか否かを立証するための方法として仮説検定の手順を整備した．検定したい仮説および対立する仮説を考え，2 種類の過誤の確率を導入し，予め定められた検定統計量の棄却域に基づいて，仮説を棄却／受容する判断を行うとした．さらに，二つの仮説がともに単純仮説の場合は尤度比検定が検出力を最大にすることを示した．

３．有意性検定と仮説検定の相違と折衷

Fisher の有意性検定は，１組のみの標本から有意か否かの推論を導き出そうとするものであり，仮説の受容という考えはない．また，過誤の確率や検出力の概念もない．他方，Neyman–Pearson の頻度流による仮説検定は，仮説の棄却／受容という二者択一の態度をとり，標本抽出と推論を無限回繰り返すという想定のもとで定義される過誤の確率を制御する（石田, 1960）．Fisher は Neyman–Pearson の仮説検定の着想に異を唱え続けた (Salsburg, 2001)．

現代一般に用いられている仮説検定の手順は両者の着想を折衷したものである．心理学系の文献では NHST (null hypothesis statistical testing) などと呼ばれている．しかしながら，最近の BASP 誌による NHST 禁止の方針 (Trafimow & Marks, 2015) や p 値に対する ASA 誌の見解 (Wasserstein & Lazar, 2016) に見られるように，p 値や仮説検定についての論争が絶えない．

４．帰無仮説の「受容」についての誤解

仮説検定および p 値についての様々な誤解の中で特筆すべきことは帰無仮説の「受容」についてである．１組のみの標本に対して p 値が大きいとしても，帰無仮説を積極的に受容する根拠とはならない．実際には差がないとは限らず，標本の大きさが十分でなかったからかもしれない．帰無仮説のもとで，p 値（下側裾確率）は一様分布に従い，一致性をもたない点に注意したい．

独立２標本問題で，等分散性の検定の結果に基づいて平均差の検定の方法（等分散 t 検定／Welch の近似 t 検定）を選択するという手順が，統計学のテキストで散見される．しかしながら，帰無仮説を受容して等分散 t 検定を実施するという手順には誤解がある．シミュレーション結果からもそのような手順の性能がいずれの t 検定の方法をも上回るという根拠は見られない．

５．結びに代えて：どう教えればよいか

データ採取の計画段階では，科学研究の再現性の観点から，標本抽出と推論の反復を想定した Neyman–Pearson の仮説検定の枠組みが有用であろう．実際，仮説検定の枠組みは標本サイズの決定に重要な役割を果たしてきた．データが採取された後の解析では，得られたデータからの事後推論の結果として，p 値とともに効果量の区間推定値を報告するのが望ましい．

初学者にとっていきなり頻度流の仮説検定の枠組みを理解することは困難を伴うと思われる．歴史的背景を考慮し，まず有意性検定，p 値について理解させ，そのあと仮説検定の枠組みの必要性を説明するのも一案であろう．少なくとも統計の授業担当者は両者の相違を意識しておきたい．