統計的仮説検定は有効か？ *

下平英寿 (統計数理研究所)

シリーズ：統計学の現状と今後

標題の問いであるが，これはもちろん YES である．しかしながら，あるひとつの方法があらゆる場面で有効であるはずはなく，優れた方法もうまく機能しない場合がある．この意味で，標題の問いに必ずしも YES とは答えられない次のような状況がしばしば指摘される．有限次元のパラメトリックモデルによって表現される帰無仮説を，それを特殊な場合として含むようなより次元の高い対立仮説に対して検定した場合，データを十分に増やしていくといずれ帰無仮説はほとんど常に棄却されてしまう．確率モデルはあくまでも現実を数学的に表現するひとつの手段にすぎないから，帰無仮説が厳密に成り立つとは考えにくい場合が多い．仮に対立仮説が成り立っているとしても，そのなかで部分モデルである帰無仮説は「測度ゼロ」であり，真の分布がちょうどその上にあるのは特殊な場合であろう．したがって，どのような帰無仮説も，十分なデータによって棄却できることになる．

このような仮説検定の「問題点」は，実用上はあまり問題にされない．たとえば 2 群の平均値 μ_1, μ_2 の比較で本来の帰無仮説が $\mu_1 \leq \mu_2$ もしくは $\mu_1 \geq \mu_2$ の場合には，第 1 種の過誤を調整するために最も不利な状況 (least favorable configuration) として $\mu_1 = \mu_2$ を想定したと考えられるので，概念的な問題はない．また実際には問題がある場合でも，帰無仮説を棄却することによって支持される対立仮説の方を示したいと統計解析をするものが望んでいる場合が多いので，かえって都合が良いのかもしれない．十分に設計・管理されていない多量のデータを集めれば，示したい結果が得られるからである．このような概念上の議論とは別に，仮説検定によって判断した結果に関してそれが十分有効に機能しているという経験的な見方もできる．帰無仮説が対立仮説の特殊な場合として得られるネストの場合には，仮説検定は実用上は有効である．

*日本統計学会 「会報」 No.100, pp.12–15. 1999 年 6 月 10 日

ところが次のような「ノンネスト」の状況では，困ったことになってしまう．まず二つの仮説 H_1 と H_2 があり，どちらも他方の特殊な場合として得られないとする．Cox 検定などの方法によって H_1 を帰無仮説，H_2 を対立仮説とする検定を行い，さらにこれらを入れ替えて，H_2 を帰無仮説，H_1 を対立仮説とする検定を行う．すると，十分なデータによって，H_1 と H_2 の両方が互いに他に対して棄却されてしまう．これは次のようにも解釈できる．H_1 と H_2 を同時に含むようなより一般的な仮説 H_X で，なるべくパラメタ次元の小さいものを考える．H_1 を H_2 に対して検定するひとつの方法は，H_1 を H_X に対して検定することであり，十分なデータによって帰無仮説 H_1 は棄却される．同様に，H_2 も H_X に対して棄却され，結果として支持されるのは H_X ということになる．このより包括的 (comprehensive) な仮説 H_X に合理的な意味が見いだせる場合にはこれで良いのだが，必ずしもそうとは限らず，このような仮説検定は有効に機能しないことになる．

わたしがこの具体例に遭遇したのは，統計数理研究所の長谷川政美先生による脊椎動物の分子系統樹の推定に関わったときである (Cao ら 1998, Mol. Biol. Evol. **15**:1637–1646)．5 群 22 種 (四足動物 15 種，肺魚，シーラカンス，硬骨魚類 4 種，ヤツメウナギ) の系統関係を調べるために，これらのミトコンドリア DNA にコードされたアミノ酸シーケンスをデータとする統計解析を行った．データ長はアミノ酸にして 3274×22 種である．進化に沿ったアミノ酸の変異過程はある種のマルコフモデルで表現され，置換モデルと呼ばれる．置換モデルを同定し，種が分化した順序を表す系統樹の形，すなわちラベル付きグラフとしての「トポロジ」を仮に与えると，ひとつのパラメトリックモデルが得られる．パラメタには，種分化の間隔を表す系統樹の各枝の長さと，置換モデルにおけるアミノ酸の変異の偏り等があり，これらは最尤法によって推定する．このモデルはサイクルのないグラフィカルモデルともみなせるが，グラフの各枝における条件付き確率が置換モデルで表現される点と，データがグラフの葉に対してしか与えられず，他のノード（＝過去の生物）はミッシングになっている点が特徴的である．

5 群の生物種の無根系統樹トポロジは 15 個あり，これらに対応する 15 個の仮説 $H_1, H_2, \ldots, H_{15}$ を互いに他に対して検定し，棄却できなかった仮説のどれかが真のものであると考えることにする．ところが，脊椎動物のデータにこの検定を行ったところ，すべてのトポロジが棄却されてしまった．相対的に最も支持されたトポロジでさえ，自由度 8 のカイ二乗分

布に近似的に従う統計量が124であった．歴史的な事実として，どれかひとつの系統樹が正しいはずである．互いに矛盾するいくつかの系統樹を合成して得られる H_X は，ミトコンドリアの場合には生物学的に受け入れられない．(なお，核遺伝子では組み替えにより H_X が妥当である場合もある.） したがって，置換モデルの特定に関する誤り (misspecification) が検出されてしまったと解釈するのが妥当である．より良い置換モデルの考案は分子系統学の重要なテーマであり，これまでに様々なものが提案されている．入手できるデータは増え続けており，それに応じてより精密なモデルを開発する必要がある．統計解析を「正しく」おこなうには，十分に現実を表現するモデルを開発してからトポロジの検定をするべきかもしれないが，実際にはなんらかの結果をすぐに出すことが要請される．ノンパラ等のロバストな手法も考えられるが，系統樹解析についてはまだ十分に研究されてはいない．このような状況で，分子系統学の分野では以下のような方法が用いられている．

仮説 $H_1, H_2, \ldots, H_{15}$ における最大対数尤度を $L_1, L_2, \ldots, L_{15}$ とする．最尤法の考え方を仮説の選択にも適用し，L_k を最大にするような仮説 $H_{\hat{k}}$ を選ぶことにする．ただし $L_{\hat{k}} = \max_k L_k$ である．系統樹の場合には，各仮説（＝モデル）のパラメタ数が等しいので，この「最尤法」は情報量規準によるモデル選択と等価である．この方法だと必ず仮説をひとつ選ぶので，仮説検定のようにすべてが棄却されてしまうことはなく，一応の結果を出せる．このようにして決めた $H_{\hat{k}}$ は仮説の「点推定」に相当し確率的なバラツキがあるから，仮説検定の p-値のような選択の信頼性を客観的に示す指標が必要である．このために提案されている方法のうち，次の３つのものについて紹介する： (i) ブートストラップ確率， (ii) ベイズの事後確率，(iii) 最大対数尤度の差の有意性検定の p-値．

(i) データ X からリサンプリングによって多数の複製 $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_N$ を生成する．$\tilde{X}_i$ における仮説 H_k の最大対数尤度を $L_k(\tilde{X}_i)$ と書く．N 個の複製のうち $L_k(\tilde{X}_i) = \max_j L_j(\tilde{X}_i)$ となる回数を N で割って，仮説 H_k のブートストラップ確率 BP_k とする．定義より $\sum_k \mathrm{BP}_k = 1$ であり，BP_k が大きいほど仮説 H_k の信頼性が高いと解釈する．この方法は非常に有用であり，手続きも簡単明瞭であるが，結果の解釈に困難がある．しかし，仮説検定の立場からの解釈と補正が Efron ら (1996, Proc. Natl. Acad. Sci. **93**:13429–34) で論じられている．

(ii) 仮説 H_k の適当な事前確率に $\exp(L_k)$ を掛けたものは，近似的にベイズの事後確率に比例する．ところが，系統樹などノンネストの仮説を扱う

ときは，仮説間の L_k の差のバラツキが大きく，$H_{\hat{k}}$ の事後確率がほとんど 1 になってしまうことが珍しくない．たとえば，真の分布に関して同程度に支持される二つの仮説から，バラツキのせいで $L_1 - L_2 = 10$ のような結果を得ることはよくあるが，ベイズファクタとして $\exp(L_1)/\exp(L_2) = 5 \times 10^5$ のように極端な数値を得てしまうので，実用上の解釈に困難がある．

(iii) 観測したデータに関して $L_1(X) - L_2(X) > 0$ だった場合，H_1 の方が H_2 よりも支持されていると考えられるが，この差が確率的なバラツキの結果である可能性もある．そこで,「二つの仮説を等しく支持する帰無仮説」を想定し，そこから X の複製 $\tilde{X}_1^*, \tilde{X}_2^*, \ldots, \tilde{X}_N^*$ を生成する．そして $L_1(\tilde{X}_i^*) - L_2(\tilde{X}_i^*)$ が $L_1(X) - L_2(X)$ より大きくなった回数を N で割った値を p_2 とする．これが有意水準より小さければ観測した差が有意であったと判断する．系統樹の場合には仮説 $H_1, \ldots, H_{15}$ があるので，これらの仮説を等しく支持する帰無仮説 H^* を想定してそこから X の複製を生成する．観測した統計量 $T_k(X) = \max_{j \neq k} L_j(X) - L_k(X)$ よりも，複製の $T_k(\tilde{X}_i^*)$ が大きくなった回数を N で割った値を p_k とする．この p-値が有意水準より小さくなる H_k は棄却される．この方法はモデル選択における仮説検定であり，論理がやや複雑なので，さらに説明を要する．

まず，この方法 (iii) は Cox 検定と次のような関係がある．仮説 H_k のモデルのパラメタを最尤推定し，これが指定する分布から生成した X の複製を $\tilde{X}_1^{(k)}, \tilde{X}_2^{(k)}, \ldots, \tilde{X}_N^{(k)}$ とする．$\tilde{X}_i^*$ の代わりに $\tilde{X}_i^{(k)}$ を用いて計算した p_k は，パラメトリックブートストラップ版 Cox 検定の p-値である．したがってモデル選択の検定と Cox 検定との違いは，帰無仮説の違いである．

上で述べた方法を明確にするには，仮説が支持される度合いを定量化する必要がある．ここでは $E(L_k(X))$ が大きいほど仮説 H_k は真の分布に関して支持されていると考える．ただし，$E(\cdot)$ はデータ X の従う真の分布に関する期待値である．$E(L_1), E(L_2), \ldots, E(L_{15})$ の最大値を $E(L_{k^*})$ とすると，k^* は未知で真の分布に依存する．そして $k = k^*$ (タイを考慮すると $E(L_k) = E(L_{k^*})$) であることを仮説 H_k^* と定義する．これは H_k が他の仮説に比べて最も支持されるような真の分布の集合ともみなせる．Kullback-Leibler 情報量の非負性より，もし H_k が真であれば H_k^* であることが示せるが，その逆は一般には言えない．厳密な意味ではすべての H_k が間違っているが，それでも $H_{k^*}^*$ は真である．なお一般には $E(L_k(X)) - \dim H_k$ や $E(L_k(X)) - \frac{1}{2}\dim H_k$ などのような次元の調整が必要であろうが，ここでは各 H_k の次元が等しい場合を考えている．

モデル選択の検定では，$H_1^*, H_2^*, \ldots, H_{15}^*$ を互いに他に対して検定す

る．第 1 種の過誤の調整のために最も不利な状況として帰無仮説 $H^* = H_1^* \cap H_2^* \cap \cdots \cap H_{15}^*$ を想定し，各仮説 H_k^* の p-値を計算する．これは $L_1, L_2, \ldots, L_{15}$ への多重比較法の適用にほかならず，下平 (1993, 統計数理 **41**:131–147)，Shimodaira (1998, AISM **50**:1–13) では近似計算によって k^* の信頼集合 ($\hat{k}$ は k^* の点推定) が与えられている．この方法の系統樹解析への応用例は下平 (1999, 統計数理 **47**, in press)，Shimodaira and Hasegawa (1999, Mol. Biol. Evol., in press) にある．また，比較する仮説が二つだけの場合については，Efron (1984, JASA **79**:791–803)，Linhart (1988, South African Statist. J. **22**:153–161)，Kishino and Hasegawa (1989, J. Mol. Evol. **29**:170–179)，Vuong (1989, Econometrica **57**:307–333)，Shimodaira (1997, AISM **49**:395–410) などで議論されている．

尤度比検定や Cox 検定などによる仮説検定はどの仮説 H_k が「正しい」かを問うのに対し，ここで述べたモデル選択の検定は，どの仮説 H_k が相対的に「良い」かを問うものである．どちらかの方法が他より優れているということはなく，両者は異なる問題である．前者については方法論としてほぼ確立しているが，後者についてはまだ十分な議論がなされていないように思える．しかしながら，後者の方法も一種の仮説検定であることには変らず，検定という考え方が広く適用できることを示している．

著者紹介： 下平英寿（しもだいら　ひでとし）．

1990 年東京大学工学部計数工学科卒業，1995 年東京大学大学院博士課程修了，博士（工学）．日本学術振興会特別研究員を経て，1996 年より統計数理研究所予測制御研究系助手，現在に至る．ニューラルネット，統計的モデル選択の諸問題，およびその分子生物学などへの応用の研究に従事．日本統計学会，応用統計学会，ASA などの会員．