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15回　無計画な解析における問題

坂巻（サカマキ）　顕太郎（ケンタロウ）*

Ⅰ　はじめに

アメリカ統計協会（American Statistical Association, ASA）が統計的有意性とｐ値に関する声明[1]を出した背景の１つに，検定（解析）結果に基づいて選択的に結果を報告するという問題がある。簡単には，ある１つのデータに対して複数の解析を実施し，ｐ値が有意水準を下回った結果（またはそれに類する結果）のみを報告する問題のことで，結果が根本的に解釈不能になるという問題が生じる。このような「結果のいいとこ取り」のことを，“cherry-picking”，“data dredging”，“significance chasing”，“significance questing”，“selective inference”，“p-hacking”などという。例えば，新たな降圧薬の治療効果を評価するために，収縮期血圧，拡張期血圧，脈圧，平均血圧の４つの評価項目に対して標準治療との群間比較を行うことを考える。簡単のため，治療効果は全く無い，それぞれの評価項目は関連しない（独立である），有意水準５％の検定を用いた群間比較を行う，という状況を考える。このとき，各評価項目の群間比較に対し，少なくとも１つ以上の検定で有意差がつく確率は，

$1-0.95^4=1-0.815=0.185(18.5\%)$

となる。これは，18.5％の確率で，「実際には治療効果が無いにもかかわらず，結果のいいとこ取りにより，治療効果が有るようにみえる結果が報告されてしまう」ということを意味する。この他にも，「ある10例の実験データで有意差がつかなかったので，別の10例を集めて検定を行い，有意差がついた10例だけの結果を報告する」「あるデータで有意差がつかなかった理由をサンプルサイズが小さいことと考え，さらにサンプルを追加したデータで検定を行い，有意差が出たデータでしか検定を行わなかったかのように報告する」などの「結果のいいとこ取り」が行われている。

「結果のいいとこ取り」はHARKing（Hypothesizing After the Results are Known：結果がわかった後に仮説を作ること）[2]とも関連している。Kerr[2]はHARKingのことを“presenting a post hoc hypothesis（i.e., one based on or informed by one's results）in one's research report as if it were, in fact, an a priori hypotheses”と述べている。例えば，レジストリデータなどの既に様々な変数が入力されているデータで解析を行い，有意差が出た結果に対し，後付けで研究仮説を作り，論文報告することがHARKingとなる。このような研究が横行した結果，再現性の危機（reproducibility crisis）が指摘されるようになり，検定やｐ値の誤用の問題が指摘されるようになった。

「結果のいいとこ取り」を防ぐには，臨床試験に関するガイドラインの１つであるICH E8（R1）（改訂された臨床試験の一般指針）[4]で指摘されているように，試験実施計画書や統計解析計画書の事前規定が重要である。この際，解析目的と解析方法における仮定を確認しておくことが必要である。例えば，収縮期血圧などの連続変数のアウトカムの平均の群間比較が目的で解析にｔ検定を用いる場合，各群のアウトカムが分散の等しい正規分布に従うとｔ検定では仮定していることを確認しておく必要がある。JAMAがまとめた統計解析計画書に関するガイドライン[5]には，“27c methods used for assumptions to be checked for statistical methods”，“27d details of alternative methods to be used if distributional assumptions do not hold, eg, normality, proportional hazards, etc”とあるように，仮定の確認や仮定が満たされな

＊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