ライティング徒然草

エグゼクティブ・アドバイザー　林 健一

第 13 回 消えていった仮説検定 (3)

本シリーズの最後として，かつては用量反応関係の評価に用いられたものの，現在ではほとんど目にしなくなった統計手法を取り上げる。具体的には，ICH E4 ガイドラインの国内向け通知「新医薬品の承認に必要な用量―反応関係の検討のための指針（平成 6 年）」の発出後の統計手法の変遷をまとめ，私なりの考察を試みることにする。

まず，本通知の発出当時によく用いられた手法は Tukey，Dunnett，Williams の多重比較である。これらは，3 群以上の試験治療群を設定したランダム化並行群間比較試験（多群試験）を実施した際，比較する群間に統計学的有意差があるかどうかを調べるための手法である。この他には，Jonckheere（ヨンキー）の検定，Cochran-Armitage の検定，最大対比法などもよく用いられた。こちらは用量反応関係の形状を評価するための手法である。すなわち，各群の反応を総合して，反応が用量に応じて上昇（または低下）するのか，それとも，ある用量で頭打ちになるのかといったことを調べるための手法で，群間比較を目的としたものではなかった。

しかし，最近承認された品目の申請資料でこれらを目にする機会は減っている。その主な理由は 2 つあり，1 つは開発品目の変化に伴う試験デザインの変化，もう 1 つは多重性（第 1 種の過誤）の調整に関する考え方の変化によってもたらされたと考えている。

まず，1990 年代は，用量反応関係を評価する目的で多群試験を実施することが多かった。被験薬の高用量群・中用量群・低用量群にプラセボ群を加えた 4 群比較試験がその典型例である。こうしたデザインは固定用量を投与する品目に適しており，降圧薬・血清脂質低下薬・血糖降下薬などの臨床試験でよく用いられた。しかし，最近では抗悪性腫瘍薬の開発に注力する製薬企業が増えている。こうした薬剤を開発する場合には，第 I 相試験で推奨用量を定めた後，第 II 相試験以降では推奨用量のみを用いて腫瘍縮小効果や延命効果を評価する。このため，多群試験を実施する機会が減り，冒頭で紹介した統計手法の出番もなくなったのである（注：ここに記載したのは化学療法薬の開発方法であり，分子標的薬では異なる開発方法をとることが多い）。

次に，多重性の調整に関する考え方も変化してきた。もともと Tukey や Dunnett の多重比較は「個々の群間比較では（2 群比較よりも）統計学的有意差が出にくくなるよう，比較する群の数に応じて第 1 種の過誤を一律に調整する」という考えに基づいて設計されたものである。

この場合，特定の群間比較で有意差が出やすくなることはない。たとえば，先の 4 群比較試験に Dunnett の検定を用いると，「高用量群対プラセボ群」「中用量群対プラセボ群」「低用量群対プラセボ群」のすべてで一律に有意差が出にくくなる。

しかし，この考え方は製薬企業の要望とは合致しない。たとえば，用量反応試験の結果に基づいて当該品目の開発を進めるか否かを判断するケースを考えてみる。この場合，多くの企業は「低用量群とプラセボ群との比較では有意差が出なくてもよいが，高用量群とプラセボ群との比較で有意差が出なければ，開発を終結する」といった判断基準を設けるはずである。すなわち，群間比較のなかには企業が重視するものとそうでないものとがあり，一律に多重性を調整する手法はこの実情にそぐわない。

そこで登場したのが「検証する仮説に優先順位をつける」という考えである。具体例を挙げたほうが理解しやすいと思うので，引き続き 4 群比較試験を題材にして，以下のように仮説に優先順位を設けることにする。

仮説 1：高用量群の主要評価項目 X の平均値はプラセボ群よりも大きい。

仮説 2：中用量群の主要評価項目 X の平均値はプラセボ群よりも大きい。

仮説 3：低用量群の主要評価項目 X の平均値はプラセボ群よりも大きい。

この場合，仮説 1 が検証できた場合にのみ仮説 2 を検証し，仮説 1 と 2 の両方が検証できた場合にのみ仮説 3 を検証するといった手順で仮説検定を実施すれば，個々の群間比較の第 1 種の過誤を調整する必要はない。すなわち，2 群比較のための統計手法を用いて，有意水準を両側 5%としたまま検定を実施してよい。もちろん，仮説 1 で有意差が得られなければ，仮説 2 の検定に進むことはできないが，高用量群の有効性が明確でなければ，企業は開発を終結するのであるから，以降の検定結果がどうなるかは問題にならないであろう。

この手順は，「仮説 1 が検証できた場合には有意水準が消費されず，次の仮説の検証時に有意水準を再利用できる」という概念を反映したもので，閉手順と呼ばれる。この概念自体は以前から知られていたが，用量反応関係評価の指針が通知された当時は，仮説に優先順位をつけることがそれほど普及していなかった。このため，用量反応関係評価に閉手順が広く用いられるようになるのは，しばらくたってからである。

さらに時代が進むと，「有意水準は分割できる」という概念も普及した。この概念の代表例は，「試験全体の第 1 種の過誤を両側 5%（0.05）として n 回の検定を行う場合には，各検定の有意水準を 0.05/n とする」というものである。これは Bonferroni の手法と呼ばれ，多重性を簡単に調整できるという利点がある反面，有意水準を必要以上に小さくするという欠点を有

していた。このため，Bonferroni 法自体が使用されることはあまりなかったが，閉手順と組み合わせることによって欠点を改善する手法が提案され，これが臨床試験の統計解析に使用されるようになったのである。

たとえば，先の 4 群比較試験で設定した 3 つの仮説に対して検定を実施し，小さな P 値から順に並べることにする。この場合，最も小さな P 値に対しては，Bonferroni 法と同様に有意水準を 0.05/3 として検定を実施する。ただし，この検定が有意になれば，残る仮説は 2 つになるので，閉手順の考えを応用して有意水準を再利用し，次に小さな P 値に対しては有意水準を 0.05/2 とする。これが Holm 法の基本原理であり，この他にも，大きな P 値から順に検定する Hochberg 法などが提案されている。

現在，Holm 法や Hochberg 法は，用量反応関係の評価だけでなく，複数のアウトカムに対する有効性を評価する場合にも用いられている。さらに，3 つ以上の試験治療群を設けるだけでなく，アウトカムも複数設けるといった複雑な試験デザインに対しては，ゲートキーピングと呼ばれる手法も用いられるようになっている。ここまでくると，「用量反応関係の評価に用いる統計手法の変遷をたどる」という当初の目的から逸脱するため，本コラムはここで終わりにするが，いずれにしても臨床試験の統計解析に用いられる手法は大きく変化した。この変化は今後も止まりそうにない。

もう戻らない
金輪際 後悔はしない

（Who-ya Extended 作詞「VIVID VICE」）